Spécificités de l'Analyse Statistique Implicative par rapport à d'autres mesures de qualité de règles d'association[1]

Especificidades da Análise Estatística Implicativa em relação à outras medidas de qualidade de regras de associação

RÉGIS GRAS[2]

RAPHAËL COUTURIER[3]

Resumo

Existem várias medidas de qualidade de uma regra de associação implicativa, analisando a semântica que guiou suas escolhas epistemológicas, percebe-se que poucas dessas medidas apoiam-se na estatística. Apresentamos aqui algumas propriedades suscetíveis de dar um significado aos índices que permitem quantificar a qualidade da associação não simétrica entre variáveis. Nesta ocasião, explicitamos, justificando nosso propósito, as diferentes escolhas que fizemos na A.S.I. para medir a qualidade das implicações entre variáveis binárias ou não e os comparamos com outras escolhas. Confrontamos as propriedades previamente enunciadas. Algumas simulações permitem ilustrar graficamente a diferença de comportamento entre alguns índices clássicas tendo em conta essas propriedades. Expomos também como a análise numérica e gráfica do conjunto de várias regras obtidas a partis de um corpus de dados conduz a uma ou várias estruturas emergentes que nosso método não simétrico, A.S.I. conceitualiza.

Palavras-chave. *Medida de qualidade. Regra de associação não simétrica. Probabilidade condicional. Implicação estatística. Implicação entrópica.*

Résumé

De nombreuses mesures de qualité d'une règle d'association implicative existent mais peu d'entre elles se fondent sur des bases statistiques et en faisant état de la sémantique qui a guidé leurs choix épistémologiques. Nous présentons ici quelques propriétés susceptibles de donner un sens aux indices servant à quantifier la qualité de l'association non symétrique entre variables. A cette occasion, nous explicitons, en les justifiant, les différents choix que nous avons faits en A.S.I. pour mesurer la qualité des implications

---

1 Cet article reprend quelques éléments du chapitre présenté sous le titre : « Gras R., Couturier R.., Blanchard J., Briand H., Kuntz P., Peter P., [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, Cepaduès –Editions, p 3-32, I.S.B.N. 18/08/20132.85428.646.4 ». Mais d'une part, il réactualise la réflexion épistémologique sur la Qualité des règles d'association, d'autre part, il met plus l'accent sur les indices définis par d'autres auteurs que ceux de l'A.S.I. et la notion de structure dérivée en A.S.I.. Il a, de plus, été présenté lors de la 5ème rencontre internationale sur l'A.S.I. (Palerme, 4 au 7 novembre 2010 en conférence introductive.

2 École Polytechnique de l'Université de Nantes, Équipe Connaissance et Décision, Laboratoire d'Informatique de Nantes-Atlantique (LINA), UMR 6241, regisgra@club-internet.fr, http://math.unipa.it/~grim/homegras_03.htm

3 Laboratoire d'Informatique de Franche Comté (LIFC), IUT Belfort-Montbéliard, Belfort, raphael.couturier@univ-fcomte.fr

entre variables binaires ou non binaires et les comparons à d'autres choix. Nous les confrontons aux propriétés énoncées préalablement. Quelques simulations permettent d'illustrer graphiquement la différence de comportement entre certains indices classiques au regard de ces propriétés. Nous exposons également comment l'analyse numérique et graphique de l'ensemble foisonnant des règles obtenues à partir d'un corpus de données conduit à une ou des structures émergentes que notre méthode systémique, l'A.S.I., conceptualise.
Mots-clés: *Mesure de qualité. Règles d'association non symétriques, Probabilité conditionnelle, Implication statistique. Implication entropique.*

Préambule

En psychologie cognitive, on considère que les connaissances opératoires et décisionnelles de l'homme soient principalement constituées et activées selon deux composantes: celle des faits et celles des règles entre les faits ou, à un niveau supérieur, entre des règles elles-mêmes. Ce serait ses propres apprentissages, à travers son éducation, sa culture ou ses expériences personnelles, qui lui permettraient une élaboration progressive des connaissances, même si des régressions, des remises en cause venaient continûment lui assurer un équilibre fonctionnel.

Sur ces bases et selon le point de vue de l'intelligence artificielle, extraire des « connaissances » à partir des données humaines, économiques, etc., c'est tenter d'observer et de nommer des régularités, des invariants, de les analyser pour en comprendre la signification afin d'essayer d'anticiper inductivement le nécessaire à partir des propriétés du contingent. Cela consiste donc à effectuer, par un traitement automatisé, des analyses interprétatives aux plans cognitif et psychologique sur la base d'un ensemble de faits, d'associations entre faits, mais également, « cerise sur le gâteau », de concepts et de structures émergées.

Parmi les associations, les règles de concomitance (« a et b apparaissent en même temps ») qui sont essentiellement *symétriques,* au même titre que la similarité, diffèrent alors des règles implicatives (« si a alors b » ou « cause → effet ») qui sont *dissymétriques*, comme est l'implication (non nécessairement causale). Or les règles, quelles qu'elles soient, se constituent de façon stable dès lors que le nombre de succès, quant à leur qualité anticipatrice ou prédictive, atteint un certain niveau de confiance en deçà duquel elles

seront susceptibles d'être plus imprudemment mises en œuvre4. En revanche, ce niveau étant atteint ou dépassé, l'économie individuelle le fera résister à l'abandon de l'embryon de règle ou, peut-être, à sa simple critique. En effet, il est coûteux, relativement à cette économie, de substituer à la règle initiale une autre règle lors de l'apparition d'un seul contre-exemple du fait qu'elle aurait été confortée par un nombre important de réussites. Un second contre-exemple (ou plus selon la qualité de robustesse du niveau de confiance subjectif en la règle) sera peut-être nécessaire à un réajustement, voire à sa mise à l'écart, ne serait-ce qu'en raison de l'éventuel danger de conserver une règle devenue inepte ou obsolète. Laurent Fleury, dans sa thèse (Fleury, 1996), cite avec pertinence l'exemple de la règle : « toutes les Ferrari sont rouges ». Cette règle, plutôt cette quasi-règle, de type implicatif, très robuste, ne sera pas abandonnée lors de l'observation d'un seul ou de deux contre-exemples. D'autant qu'elle ne manquerait pas d'être rapidement re-confortée !

Ainsi, à l'opposé de ce qui est légitime en mathématiques, en sciences « dures », où toute règle (un théorème ou une équivalence) ne souffre pas d'exception, où le déterminisme est total, où la causalité peut être invoquée, où la logique formelle est strictement respectée lui conférant a priori une prédictibilité maximale, indubitable, les règles en sciences humaines, plus généralement en sciences dites « molles », sont acceptables et donc opératoires tant que le nombre de contre-exemples restera « supportable » par la fréquence des situations où elles auront été ou seront efficaces et prédictives. On voit donc que la nature de ces règles est telle qu'elle n'exige pas le déterminisme absolu, mais ceci sans oblitérer une certaine causalité et la plausibilité d'une prédictibilité. En effet, les invariants s'enkystent au sein du désordre apparent, la nécessité niche et prospère au sein

[4] On fera, dans toute la suite, abstraction des interactions circulaires possibles de type cause a → effet b, cause b → effet c, cause c → effet a, en figeant la base de données sur un de ses états.

de la contingence. Il nous faut connaître. Et « *Connaître, c'est connaître par les causes* » affirme J. d'Ormesson (2010), dans « C'est une chose étrange à la fin que le monde »…

Le problème, en analyse des données, est alors d'établir un ou des critères, relativement consensuels, pour définir une mesure aux valeurs ajustables au niveau d'exigence de l'usager de la règle. Qu'elle soit établie sur des bases statistiques (par exemple, la quantification d'énoncés tels que « *a* apparaît généralement en même temps que *b*» ou « si *a* alors généralement *b*») a tout lieu de ne pas surprendre. Qu'elle doive posséder une propriété de résistance au bruit (faiblesse du ou des premiers contre-exemples, fortuits ou non) peut également paraître naturel, conforme au sens « économique » évoqué plus haut. Qu'elle s'affaiblisse si les contre-exemples se répètent, semble aussi guider le choix dans le modèle de mesure recherché. Cependant, comme on aurait pu le croire, les modélisations associées où intervient cette mesure, ne sont pas celles des tests d'hypothèse nulle où celle-ci serait rejetée dès qu'un seuil d'invraisemblance serait atteint. Mais également, la philosophie est différente comme l'énonce de manière antithétique I.-C. Lerman (Lerman, 1992) :

> ...la philosophie de l'analyse de données est en quelque sorte opposée à celle des tests d'indépendance. Pour cette dernière, on a relativement à l'existence d'un lien : ' FAUX JUSQU'À PREUVE DU CONTRAIRE' alors que pour l'optique de l'analyse de données, on a 'VRAI JUSQU'À PREUVE DU CONTRAIRE.

Autrement dit, ce qui importe dans l'analyse de données, ce sont les mesures qui quantifient les associations et la relation d'ordre ou de pré ordre entre elles.

Se pose alors le problème de *qualité d'une règle*, voire la *mesure de sa qualité* ou plus encore, de la *qualité de la mesure de qualité de règles*[5], c'est–à-dire selon nous, le problème qui passe par l'étude de la satisfaction optimale, ou dans certains cas, locale, d'un certain nombre de critères qualitatifs et quantitatifs, comme ceux évoqués ci-dessus. Des mesures de qualité de

5 car le crédit à accorder à une règle ne peut être pertinent qu'à condition que la mesure qui l'évalue respecte des propriétés de fidélité et d'adéquation sémantique et analytique. Cependant, nous désignerons de façon identique ces deux concepts pourtant distincts.

règles existent nombreuses dans la littérature et les usages. Nous-mêmes en avons définie, l'A.S.I. (Gras, 1979). Ainsi la qualité d'une règle qui se pose en concept scientifique, « pour éminemment objective que soit sa portée, implique une activité subjective de la pensée » (Sève, 2005, p.108). Elle est même de nature dialectique (logique de la contradiction « dépassable »). En voici une illustration :

- sur un ensemble de 10 individus, les attributs *a* et *b* sont vérifiés respectivement 6 et 8 fois, sans contre-exemples à la règle « si *a* alors *b* ». Celle-ci est donc logiquement acceptable,
- sur un ensemble de 1000 individus, les mêmes attributs sont vérifiés respectivement 600 et 800 fois, avec deux seuls contre-exemples à la règle. La règle logique n'est plus acceptable, mais on pourrait pragmatiquement y croire et l'adopter.

A laquelle accorderiez-vous la meilleure qualité si vous utilisiez comme critère la propriété ordinale du nombre de contre-exemples observés ou encore du nombre relatif de contre-exemples à b sachant a (emploi de la propriété de la fréquence conditionnelle) ? Dans le premier cas, la règle est stricte mais la confiance en elle est faible car elle est insuffisamment observée. Dans le second cas, c'est le contraire. Cette contradiction relative à la validité de la règle doit être effectivement et dialectiquement dépassée. On perçoit immédiatement que la mesure de qualité devrait passer sous les fourches de la statistique.

Nous nous proposons dans ce texte d'établir un inventaire, nécessairement subjectif mais visant, peut-être, l'exhaustivité, de propriétés liées à une certaine qualité de mesure, et, de façon critique, de les illustrer, sans exhaustivité cette fois, par des mesures existantes, puis de les confronter à notre propre mesure, l'A.S.I., à l'aune de ces propriétés. Nous éludons volontairement, par conséquent, différentes autres questions qui seraient rattachées à ce problème de qualité de règles dans le cadre de l'extraction de connaissances à partir de données. Ainsi, sont réservées à d'autres études certaines de ces questions évoquées dans cet article.

Quelques petits détours « philosophiques » apporteront un éclairage nouveau pour aborder ici et ultérieurement ces mêmes questions.

Nous regroupons ci-dessous selon trois rubriques celles de ces questions qui nous paraissent les plus vives et selon lesquelles pourrait être envisagé le problème de la qualité :

- Quel est l'objet sur lequel porte la problématique de la qualité : les données ? le processus d'extraction des données ? les connaissances extraites ? les connaissances constituées de règles ?
- Quel est le point de vue selon lequel on se place ? celui des objectifs, celui des attentes, celui des besoins de l'utilisateur ? selon une certaine pertinence par rapport aux données traitées? quelles sont les mesures de qualité connues ? comment les comparer ? quelles comparaisons et selon quels critères ? comment aider l'utilisateur à faire un choix ? (cf. LENCA et al, 2004)
- Comment définir une mesure ? selon quels critères ? quels indices seraient à prendre en compte ? quelle mesure ? quelle modélisation ? quelle formalisation ? quel instrument ou quelle méthode de mesure ? l'appréciation de qualité doit-elle être sémantique ? statistique ? les deux ? une règle se compose-t-elle de plusieurs prémisses ? de plusieurs conclusions ? comment mesurer la cohérence entre les règles extraites ? quelle information extraire également des règles de règles ? quels en sont l'utilité ? le sens ? comment les interpréter ? quelles aides les représentations graphiques peuvent-elles apporter à la lecture de règles ? quelles nouvelles pistes pourrait-on ouvrir ?

Afin de mieux circonscrire notre travail, nous tentons, dans cet article, de donner des réponses aux questions relevant principalement de la dernière rubrique. On trouvera dans les ouvrages cités en référence des approches comparables, reprenant quelquefois celle de cet article et indiquant d'autres indices de règles d'association et d'autres propriétés de qualité. Nous en évoquerons quelques-uns au fil du texte présent et, en particulier, dans le § 4.

A la fin de cet article, nous élargissons la réflexion au-delà du problème seul des mesures de qualité. Ceci nous conduit à présenter ici en pleine cohérence les *objectifs* suivants :

- étudier différents modèles de mesure de qualité de règles, voire de métarègles non symétriques,
- *évaluer* numériquement ces qualités et, **en outre**,
- *structurer* les ensembles de règles et métarègles retenues à un seuil choisi par l'utilisateur et
- *représenter* les structures obtenues.

Introduction

La situation paradigmatique de recherche de règles se ramène à croiser des variables binaires (attributs, caractères, etc.) et des sujets ou des objets sur lesquels on observe les variables à travers des instances en nombre *n*. Puis à étudier, généralement, les relations qu'entretiennent entre elles, dans l'observation, les variables ou des conjonctions (dites items-sets) ou des disjonctions de celles-ci. Ainsi, *sur le plan opératoire*, une relation de similarité de bonne qualité, observée à partir d'une importante concomitance de deux variables *a* et *b*, permettra en toute hypothèse de préjuger la présence de *a* quand on observera *b* et réciproquement. Mais aussi, une relation d'implication, observée à partir d'une quasi-inclusion de l'ensemble des individus qui satisfont *a* parmi ceux qui satisfont *b* pourrait permettre de prédire *b* quand on observera *a*. Cependant, cette opérativité n'aura de consistance que dans la mesure où la qualité (affirmative) de la règle aura été reconnue adaptée et l'instrument qui l'évalue aura été adéquat[6] à une *sémantique* de l'implication qui reste à formaliser. Ce sont les problèmes théoriques qui nous préoccupent ici[7]. Ils vont consister à mettre en question la formule de R. Thom (THOM,

[6] dans l'hypothèse encore, que nous admettrons, où le recueil des données et celles-ci auront été de bonne qualité

[7] Comme le dit Ervin Laszlo dans « La cohérence du réel » (Gauthier-Villars, 1989) : «Il n'y a pas d'immaculée perception : nous ne voyons la réalité qu'au travers des lunettes de la théorie », p. 186.

1980) : « Ce qui limite le vrai n'est pas le faux mais l'insignifiant » car c'est justement ce *quasi* auquel nous donnerons un sens et une mesure.

Dans la littérature, on trouve le plus fréquemment 3 paramètres constitutifs de la mesure de qualité :

- le ***support*** de la règle associant les variables binaires *a* et *b* (d'occurrences respectives n_a et n_b) qui est la fréquence $\frac{n_{a\wedge b}}{n}$ d'occurrences simultanées de *a* et *b* parmi les n instances, qui servira d'indicateur aussi bien de la concomitance que de l'implication ;
- le ***coefficient de corrélation linéaire*** qui sera un des indicateurs privilégiés pour signifier principalement cette concomitance (très souvent abusivement utilisé comme révélateur de causalité);
- la fréquence conditionnelle[8] de *b* sachant *a*, F(b/a) ou $c=\frac{n_{a\wedge b}}{n_a}$, appelée aussi ***confiance c*** qui servira surtout dans la recherche de règle implicative car elle n'est pas symétrique en *a* et *b*, contrairement aux deux paramètres précédents.

Or, nous convenons aisément qu'une règle entre deux variables binaires *a* et *b* ou entre des conjonctions de variables, règle du type $a \Rightarrow b$ ou $a \Leftrightarrow b$, est dite- simple raison de bon sens- de **bonne qualité** lorsque le nombre de contre-exemples (d'occurrences $n_{a\wedge\bar{b}}$, où $\bar{b}$ désigne conventionnellement *non b*) à sa validité formelle est nul ou faible. Une difficulté apparaît alors : quand dira-t-on que le nombre non nul de contre-exemples est faible ? On peut alors adopter plusieurs indicateurs de rareté dans un ordre décroissant de « naïveté » :

- la rareté absolue des contre-exemples, i.e. un nombre de ceux-ci peu différent de zéro : 1 ou 2 ou 3…, selon un jugement subjectif, mais consensuel ;

[8] souvent dénommée à tort « probabilité conditionnelle » au lieu de « fréquence conditionnelle », expression que nous utiliserons fréquemment pour respecter la coutume

- la rareté relative des contre-exemples ou, a contrario, l'abondance relative des exemples qui satisferaient qualitativement la sémantique de l'association, celle-ci s'exprimant, selon sa nature, de la façon suivante : pour l'implication, au sein d'une population non réduite, si *a* est vérifié, *b* doit l'être *toujours ou presque toujours* ; pour l'identité ou pour l'équivalence, *a* et *b* doivent avoir *toujours ou presque toujours* les mêmes occurrences ;
- une « bonne » fréquence de *b* sachant que *a* est réalisé (et/ou de *a* sachant *b*) ;
- une combinaison de cette fréquence avec d'autres fréquences ou avec le coefficient de corrélation linéaire ;
- la réfutation d'un test d'hypothèse où l'hypothèse d'indépendance de *a* et de *b* constituerait l'hypothèse nulle ;
- la forte entropie de l'expérience où l'on comparerait les nombres d'exemples et de contre-exemples, le premier l'emportant manifestement sur le second,
- la valeur de la probabilité, dans l'hypothèse d'indépendance entre *a* et *b*, d'avoir plus de contre-exemples dans une expérience aléatoire de mêmes caractéristiques cardinales, que ceux qui ont été observés, etc. .

Il semble alors incontournable, dans ces conditions, où des valeurs numériques sont en jeu, et si l'on souhaite objectiver la qualité, que l'on doive affecter une mesure à celle-ci. Quelles propriétés devrait alors posséder cette mesure ?

Voici quelques propriétés qui nous semblent fondamentales et, semble-t-il, relevables d'un bon sens … dont nous ne contestons pas la subjectivité. Elles seront quelquefois illustrées par des graphiques, commentées dans le corps du texte. **Nous privilégierons les mesures portant sur les règles implicatives** en renvoyant le lecteur, par exemple, aux travaux voisins de I.C. Lerman (LERMAN 1970, 1981b, 1984, 1992) sur les indices de similarité, dont nous nous sommes primitivement inspirés.

Quelques propriétés pour un indice de qualité d'une règle d'association

a) Propriété P_1 : tenir compte du nombre de contre-exemples

Il semble légitime qu'une mesure de qualité doive croître avec un ou plusieurs des indicateurs ci-dessus, en particulier relativement au **nombre de contre-exemples**, les valeurs extrêmes de la mesure pouvant être atteintes respectivement dans le cas de rejet de la règle ou d'acceptation « indubitable » de celle-ci. **Autrement dit, la valeur numérique de la mesure associée à une règle doit être fonction croissante de sa qualité sémantique**. C'est le cas pour les 4 premiers exemples ci-dessous. Des valeurs positives de cette mesure semblent mieux adaptées à rendre compte de la qualité que des valeurs négatives, en raison de la relation d'ordre naturelle sur les valeurs absolues des réels.

Exemple 1 : le coefficient de corrélation de Pearson croît vers +1 lorsque les deux variables sont de plus en plus ressemblantes i.e. ne s'opposent pas ;

Exemple 2 : l'indice de Loevinger, (LOEVINGER 1947), l'aîné semble-t-il des indices d'implication de *a* sur *b*, $H(a,b) = 1 - \frac{n.n_{a\wedge\bar{b}}}{n_a n_{\bar{b}}}$, prend ses valeurs sur $]-\infty, +1]$, croît vers 1 quand le nombre de contre-exemples à l'implication , $n_{a\wedge\bar{b}}$, tend vers 0 ; il tend vers 1 (prolongement par continuité) lorsque la variable b devient de plus en plus sûre ($n_{a\wedge\bar{b}}$ est, par exemple, un infiniment petit d'ordre supérieur à celui de $n_{\bar{b}}$) et prend des valeurs négatives, en particulier lorsque les variables *a* et *b* sont indépendantes.

Exemple 3 : L'indice appelé Zhang défini dans (TERANO et al., 2000) s'appuie au contraire sur l'abondance des exemples : $z = \frac{n\, n_{a\wedge b} - n_a n_b}{max[n_{a\wedge b}, n_{\bar{b}}, n_b, n_{a\wedge\bar{b}}]}$. Cette expression se simplifie, pour $n_a \leq n_b$, en $z = \frac{n\, n_{a\wedge b} - n_a n_b}{max[n_{\bar{b}}, n_b]}$ du fait que $n_{a\wedge\bar{b}} \leq n_{\bar{b}}$ et $n_{a\wedge b} \leq n_b$. Elle croît avec le nombre d'exemples, est égale à $n_{\bar{b}}$ (resp. n_b) lorsqu'il n'y a pas de contre-exemples et que $n_b \geq n_{\bar{b}}$ (resp. $n_b < n_{\bar{b}}$).

On verra plus loin que cet indice possède certaines affinités, au niveau de la qualité, avec l'indice de base (classique) de l'A.S.I..

Exemple 4 : *L'indice de base* pour la mesure de l'implication en A.S.I. s'exprime par:

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

Et l'intensité d'implication, dite classique, qui en mesure la qualité est, pour $n_a \leq n_b$:

$$\varphi(a,b) = 1 - \sum_{s=0}^{n_{a\wedge\bar{b}}} \frac{\hat{\lambda}^s}{s!}.e^{-\hat{\lambda}} \text{ avec } \hat{\lambda} = \lambda_{estimé} = \frac{n_a.n_{\bar{b}}}{n} \leq 4$$

$$et\ \varphi(a,b) = \frac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt \quad pour \frac{n_a n_{\bar{b}}}{n} > 4 .$$

Elle exprime concrètement « l'étonnement statistique » d'observer plus de contre-exemples dans l'hypothèse d'indépendance implicative que dans la contingence, philosophie courante en statistique non paramétrique.

Exemple 5 : A l'opposé, avec la J-mesure de $a \Rightarrow b$ (GOODMAN-SMITH, 1989) définie par :

$$J(b;a) = p(a).j(b;a) = p(a).\left[p(b \mid a).\log_2\left(\frac{p(b \mid a)}{P(b)}\right) + p(\bar{b} \mid a).\log_2\left(\frac{p(\bar{b} \mid a)}{p(\bar{b})}\right)\right] =$$

$$p(a \wedge b).\log_2 \frac{p(a \wedge b)}{p(a).p(b)} + p(a \wedge \bar{b}) \log_2 \frac{p(a \wedge \bar{b})}{p(a).p(\bar{b}}$$

la valeur seule de la J-mesure ne permet pas de savoir laquelle des deux implications $a \Rightarrow b$ ou $a \Rightarrow \bar{b}$ est dominante car sa valeur maximale ne correspond pas à la meilleure qualité de la première des deux règles (voir la figure 4). De plus les valeurs extrêmes de la mesure ne sont

pas des constantes. Ce qui est aussi le cas, comme nous le verrons dans 3.4 avec l'indice de Piateski-Shapiro.

Si la propriété **P_1** est globalement satisfaite en A.S.I., en revanche dans le cas où les instances deviennent très grandes (la négativité du numérateur diminue avec 1/n), la tendance de la mesure à prendre la valeur 1 rend délicate l'affirmation d'une bonne et « étonnante » qualité de l'implication. On verra que, suite à cette observation, le passage à la forme entropique corrige ce défaut.

Exemple 6 : La mesure appelée Lift (BRIN et al., 1997) est exprimée par : $\frac{n.n_{a\wedge b}}{n_a\, n_b}$

Elle ne fait pas intervenir explicitement le nombre de contre-exemples. Mais si celui-ci est nul, le Lift vaut n/n_b .

b) Propriété P_2 : prendre en compte la dépendance implicative

La recherche de la règle d'association entre *a* et *b* peut se faire sur la base de négation de **l'indépendance et plus précisément d'absence de liaison implicative entre ces deux variables**. C'est donc cette indépendance qui est interrogée et évaluée, dans l'A.S.I., dans la mesure de qualité.

$\varphi(a,\bar{b})$ est **la probabilité gaussienne** pour que la contingence conduise à relativement moins de contre-exemples à l'implication auquel conduirait le seul hasard si les variables *a* et *b* étaient indépendantes. Par suite, elle est d'autant plus grande que $q(a,\bar{b})$ est négatif, c'est-à-dire que le nombre de contre-exemples est faible eu égard à celui qu'aurait donné le simple hasard et que n_a étant grand avoisine n_b[9](inclusion des exemples de *a* dans ceux de *b*). Ainsi en A.S.I., on dispose d'une échelle de mesure de la qualité : variant de 0 à 1, la valeur 1 correspond à une très bonne qualité de la règle.[10]

[9] L'étonnement statistique en est une métaphore

[10] Dans (GRAS et al, 1996), nous étudions la qualité de l'intensité d'implication et donnons une échelle qui mesure la pertinence de la valeur observée $\varphi(a,\bar{b})$

Elle pourrait l'être aussi, mais n'est ni explicitée ni étudiée, dans la mesure de Loevinger ni dans l'indice de Piatetski-Shapiro. Or le choix d'une telle mesure dans le cadre de l'analyse de données s'oppose à la « philosophie » des tests d'hypothèse, comme nous l'avons évoqué dans le préambule, tout en présentant, dans la définition de la mesure, des analogies avec certains tests non paramétriques. Dans bon nombre de ceux-ci, on mesure l'écart entre une distribution théorique si seul le hasard intervenait et la distribution observée (ex. test du χ^2 ou test de Kolmogoroff). Selon cette « philosophie », dans une problématique d'extraction de connaissances, supposant l'existence d'une population-mère d'où serait extrait l'échantillon de l'expérience, il serait possible, avec l'A.S.I., de construire un test où l'hypothèse nulle s'exprimerait par *« dans les conditions cardinales données, si les variables a et b étaient indépendantes, la fréquence de réalisation de contre-exemples à l'équivalence ou à l'implication, devrait être inférieure à une certaine valeur □ avec une probabilité inférieure à □ »* (ou plus simplement, « *le nombre de contre-exemples devrait être nul, sous la même hypothèse d'absence de lien »*). La règle de décision serait alors : « *si l'observation conduit à une valeur supérieure à □ on rejette l'indépendance a priori »*. Ainsi, cette attitude de réfutation n'offre que l'alternative : acceptation ou rejet à un seuil donné. Ce n'est pas celle qu'adopte l'analyste de données.

En revanche, en effet, en analyse de données, sans contrainte d'échantillonnage, la mise en place d'une mesure, comme celle qui découle de l'algorithme de la vraisemblance du lien (LERMAN 1970, 1984), utilisé originellement pour des règles implicatives par nous (GRAS, 1979), permet d'affecter une valeur numérique à chaque couple de variables. Ainsi il est possible de les organiser, de les synthétiser, de structurer leur ensemble au moyen de ces valeurs, ce que n'est pas l'objectif majeur d'un test d'hypothèse. Ce qui n'interdit pas de retenir un seuil pour représenter la structure et de considérer que, pour des cardinaux observés, plus la valeur est proche de 1, plus la liaison implicative est forte, ou pour le moins crédible. Et ceci devra, bien sûr, **s'opposer** dans ce cas, à **l'indépendance des variables** pour laquelle la

fréquence, dans le cas de l'implication par exemple, de *a* et *non b* serait égale au produit des fréquences respectives de *a* et de *non b*.

De son côté, Julien Blanchard (BLANCHARD et al, 2005a), plutôt que de retenir une mesure prenant en compte l'écart entre l'indépendance a priori et la dépendance contingente, a choisi de mesurer par un indice dit IPEE (ou Indice Probabiliste d'Ecart à l'Equilibre) la significativité de l'absence d'équi-répartition entre le nombre d'exemples de (a et b) et celui des contre-exemples (*a* et *non b*) au sein des instances de *a*. Constatant le caractère peu discriminant de cet indice pour de grandes valeurs de n, il étend cet indice, comme nous l'avons fait avec l'intensité d'implication, selon un indice entropique. On retrouve dans le Lift (ex. 6 du 3.1) cette prise en compte de la comparaison entre l'indépendance et les exemples de la règle, mais elle n'est pas probabiliste contrairement à la mesure IPEE. .

c) Propriété P_3 : respecter des comportements attendus de la mesure

La décroissance de la mesure de qualité doit respecter certaines attentes sémantiques où, de plus, interviennent les coûts respectifs de la remise en cause ou celui de la conservation d'une association non tenable:

- décroissance lente au début de l'apparition des contre-exemples car le fortuit, le bruit ou les erreurs d'observation sont peut-être la cause de leur apparition (nécessité de la non-remise en cause trop rapide d'une règle universelle). Souvenons-nous de l'exemple donné par L. Fleury et cité dans l'Introduction où une règle admise généralement et confortée par l'expérience ne serait pas réfutée par un ou deux contre-exemples car elle est consensuellement résistante,
- décroissance plus rapide dès que l'observation des contre-exemples confirme l'absence manifeste d'une règle d'association.

En conséquence, **la mesure ne doit pas varier linéairement avec le nombre de contre-exemples** car chacun d'entre eux n'a pas le même poids dans l'économie psychologique de la règle et peut naître du bruit.

Or **la linéarité** admet deux composantes épistémologiques qui coexistent[11] et que nous prenons en compte dans l'A.S.I. : *la proportionnalité* qui s'oppose à la sémantique d'une règle d'association et l'*additivité* qui s'opposerait à la recherche de structure cohérente (GRAS et al., 2009).

Exemple 1 : Par suite, la fréquence conditionnelle, fonction linéaire des contre-exemples, intéressante pour signifier la **confiance**, ne peut pas être utilisée seule, selon nous, de même que toute combinaison linéaire ou non de ce type[12], comme, par exemple, l'indice de J.G. Ganascia (GANASCIA 91) : c= 2Prob(b/a). En effet, leurs dérivées partielles par rapport à la variable « nombre de contre-exemples » est constante.

Par exemple, $\frac{\partial c}{\partial n_{a\wedge\bar{b}}} = -2\frac{1}{n_a}$ montre que la confiance décroît lorsqu'augmentent les contre-exemples mais en ne dépendant que de n_a.

Exemple 2 : Figura 1, montre, par contre, la différence de comportement, suivant les valeurs du nombre de contre-exemples, $n_{a\wedge\bar{b}}$, entre la fréquence ou probabilité conditionnelle (confiance), le coefficient de corrélation linéaire et l'intensité d'implication classique ou entropique [13](GRAS et al, 1996, 2008 et 2009).

---

[11] Citons, à ce sujet, deux extraits du chapitre de L.Sève (SEVE et al, 2005, p.58-59) :
« …le tout ne se compose de rien d'autre que de ses parties, et pourtant il présente en tant que tout des propriétés n'appartenant à aucune 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 » ;
« Tout se passe donc comme si se produisait une *génération spontanée* de propriétés du tout…C'est le paradoxe de **l'émergence**. »

[12] On a vu dans l'introduction un exemple où la confiance au sens ci-dessus ne conduit pas nécessairement à la « confiance » au sens de tous les jours.

[13] Rappelons que cette deuxième forme (implication entropique) de l'A.S.I. est basée sur un indice dit d'inclusion qui est d'autant plus fort que sont faibles les entropies conditionnelles, d'une part de b sachant a et d'autre part de non a sachant non b. C'est-à-dire que sont élevées les informations respectives de l'inclusion des exemples de a dans ceux de b (qualité de $a \Rightarrow b$) et des exemples de non b dans ceux de non a (qualité de la contraposée *non b* $\Rightarrow$ *non a*). L'intensité d'inclusion associe alors cet indice d'inclusion et l'intensité classique d'implication.

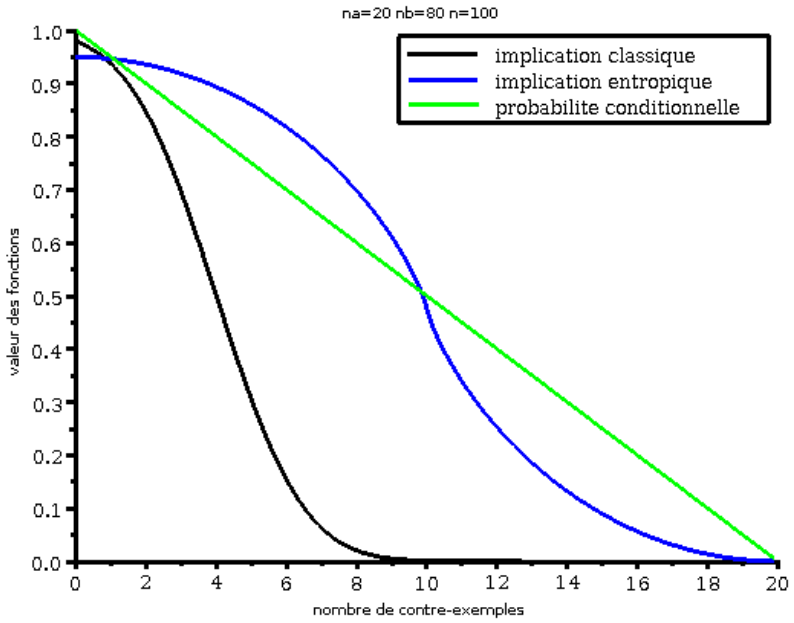
Par exemple, l'indice d'implication classique en A.S.I. varie avec le nombre de contre-exemples selon

$\frac{\partial q}{\partial n_{a\wedge\bar{b}}} = \frac{1}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}} > 0$. Ceci signifie que l'indice q croît avec ce nombre. L'intensité d'implication décroît en conséquence et d'autant plus vite que *a* ou/et *non b* deviendraient moins fréquents (*b* plus fréquent).

Ainsi, les deux formes de l'ASI satisfont cette propriété **P_2.**

Exemple 3 : A l'instar de ce que nous faisons en A.S.I., **l'indice probabiliste d'écart à l'équilibre** de Julien Blanchard (BLANCHARD J., 2005) mesure l'écart à l'équi-répartition des exemples de (*a et b*) et des contre-exemples (*a et non b*). Plus cet écart est important comparé à celui que donnerait le seul hasard dans une hypothèse d'indépendance de *a* et *b*, plus l'implication est de bonne qualité. Cet indice satisfait également **P_2.**

Figura 1. Comparaison des variations de l'implication classique, l'implication entropique, le coefficient de corrélation linéaire et la probabilité conditionnelle en fonction du nombre de contre-exemples



d) Propriété P_4 : respecter la cohérence sémantique

Toujours pour des raisons sémantiques, d'autres contraintes semblent à respecter pour des règles implicatives :

- mesurer la qualité **ne doit pas se faire en occultant les cas rares** où peuvent se nicher des « **pépites de connaissances** » (KODRATOFF, 00). La seule propriété de support des variables serait dans ce cas scotomisante car la rareté d'une prémisse l'écarterait de sa fonction potentielle de cause sur des effets possibles. **L'effet papillon** n'en est-il pas la meilleure illustration ? C'est pourtant ce qui se produit avec la confiance et les mesures dérivées ce qui nous semble susceptible de les disqualifier partiellement pour cette raison ;

- la mesure **ne doit pas conduire à des paradoxes ou des incohérences** entre la valeur de qualité et l'absence certaine d'une quelconque association ; ainsi, ne pas prendre les mêmes valeurs pour des situations manifestement différentes quant aux règles attendues.

Par exemple, nous citons en note deux situations assez paradoxales auxquelles conduit l'indice « surprise »[14] (AZE, KODRATOFF, 01) : $\frac{P(a,b)-P(a,\bar{b})}{P(b)}$ ou $\frac{n_{a\wedge b}-n_{a\wedge \bar{b}}}{n_b}$ qui, par ailleurs, présente certaines autres qualités (Figura 2);

- elle **ne devrait pas être symétrique**, c'est-à-dire que, s'il n'y a pas équivalence, la mesure associée à la règle $a \Rightarrow b$ devrait généralement différer de celle de la réciproque $b \Rightarrow a$. Ce serait, hélas le cas, si en A.S.I. on choisissait la modélisation hypergéométrique de la variable aléatoire « nombre de contre-exemples », choix qu'ont adopté certains auteurs pendant quelque temps. C'est le cas de l'indice de Piatetski-Shapiro : $S= \frac{1}{n}(\frac{n_a n_{\bar{b}\cdot}}{n} - n_{a\wedge\bar{b}})$ qui accorde la même valeur à la règle directe et à sa réciproque (voir aussi à ce sujet figure 3) ; c'est aussi le cas de la mesure Lift ce qui les discréditent sémantiquement ;

Figura 2. Comparaison de la probabilité conditionnelle, de l'implication entropique et de la « surprise »

[14] Ex 1 : pour n donné, $n_{a\wedge b}$ = 9, $n_{a\wedge \bar{b}}$ =3, $n_{\bar{a}\wedge b}$ =3, conduit à une surprise 0.5 de la même façon que si $n_{a\wedge b}$ = 12, $n_{a\wedge \bar{b}}$ =0, $n_{\bar{a}\wedge b}$ =12. Or l'inclusion de A dans B est totale dans le 2ème cas.
Ex 2 : 3 situations différentes : $n_{a\wedge b}$ =80, $n_{a\wedge \bar{b}}$ =0, $n_{\bar{a}\wedge b}$ =10, puis $n_{a\wedge b}$ =8, $n_{a\wedge \bar{b}}$ =0, $n_{\bar{a}\wedge b}$ =1, et enfin $n_{a\wedge b}$ = 85, $n_{a\wedge \bar{b}}$ =5, $n_{\bar{a}\wedge b}$ =5 conduisent à la même valeur de surprise 8/9.

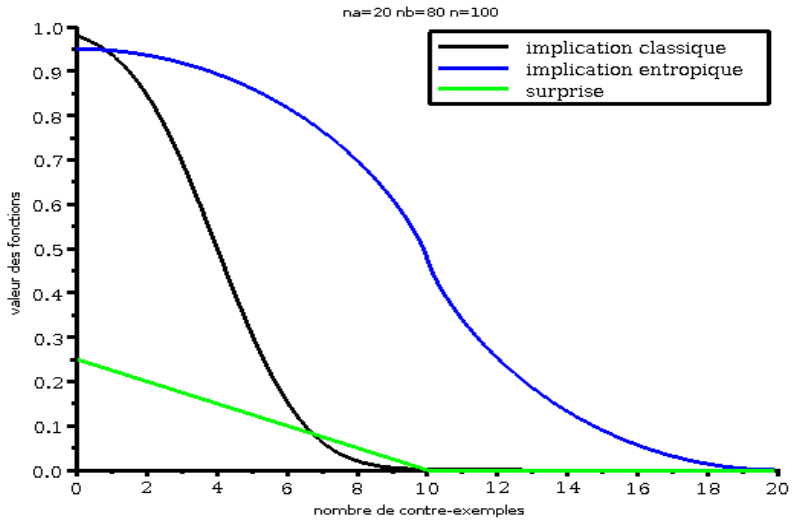
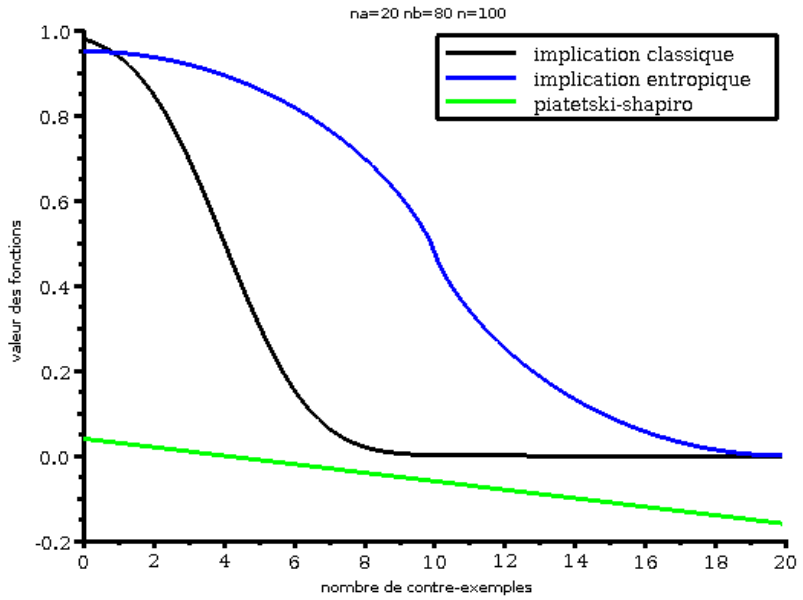


Figura 3. Comparaison de l'implication classique, de l'implication entropique et de l'indice de Piatetski-Shapiro



- elle doit être **sensible aux variations des paramètres** intervenant dans la relation implicative, n, n_a, n_b et $n_{a\wedge\bar{b}}$. Nous venons de le voir pour le nombre de contre-exemples. Nous en reparlerons pour les occurrences de b dans l'examen de la propriété P_5 et pour n dans celui de la propriété P_7. Mais notons que la « surprise » n'est pas sensible à n_a, que la confiance ne l'est ni pour n, ni pour n_b, que l'indice de M.Sebag et M. Schoenauer ne l'est pas non plus, que la J-mesure n'est pas sensible à n, **mais** que l'indice de Loevinger et l'indice d'implication le sont pour les 4 paramètres. En effet, étendant l'indice d'implication A.S.I. q(.), fonction, des 4 variables liées n, $n_{a\wedge\bar{b}}$, n_a, n_b à valeurs dans N en une fonction sur R^4 différentiable par rapport aux variables en jeu dans $q(a,\bar{b})$, sa différentielle est : $dq = \frac{\partial q}{\partial n} dn + \frac{\partial q}{\partial n_a} dn_a + \frac{\partial q}{\partial n_b} dn_b + \frac{\partial q}{\partial n_{a\wedge\bar{b}}} d n_{a\wedge\bar{b}}$;

- enfin, elle **ne doit pas conduire, de façon également paradoxale**, à affecter la même mesure aux règles $a \Rightarrow b$ et $a \Rightarrow \bar{b}$ ce qui est pourtant le cas de la J-mesure de Goodman-Smyth (Goodman et al. 1989) (cf. 3-2).

e) Propriété P_5 : assurer une décroissance de la mesure avec la trivialité

Afin que la mesure ait une fonction informative, **elle doit décroître avec la trivialité des observations**, en appelant « observation triviale » toute observation contenant une information prévisible ou nulle au sens de l'entropie de Shannon (incertitude faible ou nulle). Par exemple, si l'on examine la règle $a \Rightarrow b$, la mesure doit décroître, pour a fixée, quand la réalisation de b est presque sûre et, a fortiori, certaine. De même et en outre, dans le cas de l'équivalence, si a et b sont presque certains, voire certains, la mesure doit être faible voire nulle car elle est devenue triviale.

Exemple 1 : La fréquence conditionnelle présente le caractère d'assurer à la règle $a \Rightarrow b$ la valeur 1 de confiance maximum si b est sûre (b=1 dans toutes les instances). La règle n'est pas informative car connue.

Exemple 2 : Ce n'est pas le cas de l'intensité d'implication dont on montre (GRAS et al. 1996), par un prolongement par continuité, la convergence de l'intensité vers 0 lorsque n_b tend vers n (trivialité de *b*).

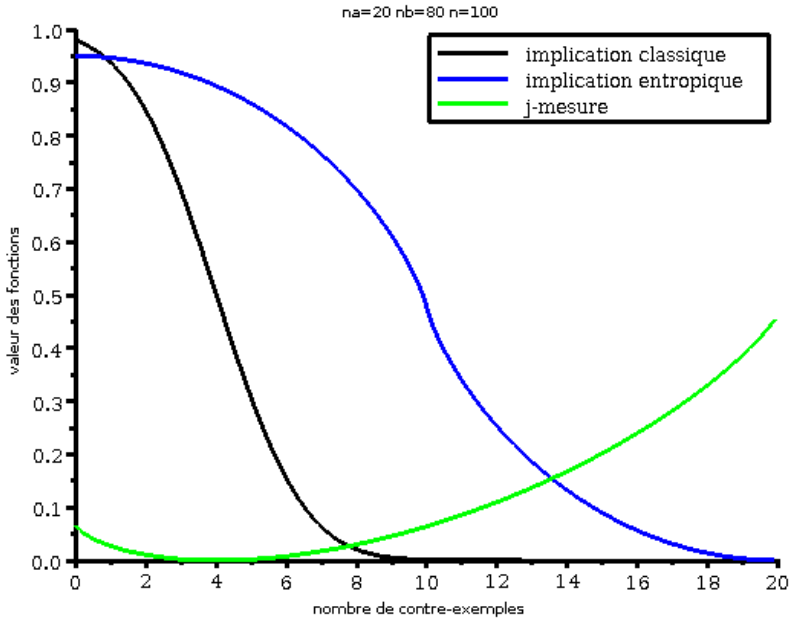
Exemple 3 : L'indice de M.Sebag et M. Schoenauer (SEBAG et al. 1991) défini par l'inégalité suivante où est un seuil d'acceptabilité de la règle $a \Rightarrow b$:

$\frac{\text{Prob}[a \wedge b] - \text{Prob}[a \wedge \bar{b}]}{\text{Prob}[a \wedge b]} \geq \alpha$ satisfait l'inégalité pour tout quand *b* est certaine.

Exemple 4 : Par contre, l'indice de Ganascia, défini dans 3.3, n'est pas maximum lorsque *b* est certaine et satisfait donc P_4.

Exemple 5 : La figure 4 montre la simulation selon le nombre $n_{a \wedge \bar{b}}$ de contre-exemples à la règle $a \Rightarrow b$ des indices de l'intensité classique d'implication en A.S.I. ou avec la confiance et la J-mesure. On constatera que la J-mesure ne prend pas la valeur 1 lorsque $n_{a \wedge \bar{b}}$ est nul, décroît vers 0 quand ce nombre augmente, pour croître à nouveau à partir d'une certaine valeur inférieure à n_a.

Figura 4. Comparaison de l'implication classique, de l'implication entropique et de la J-mesure

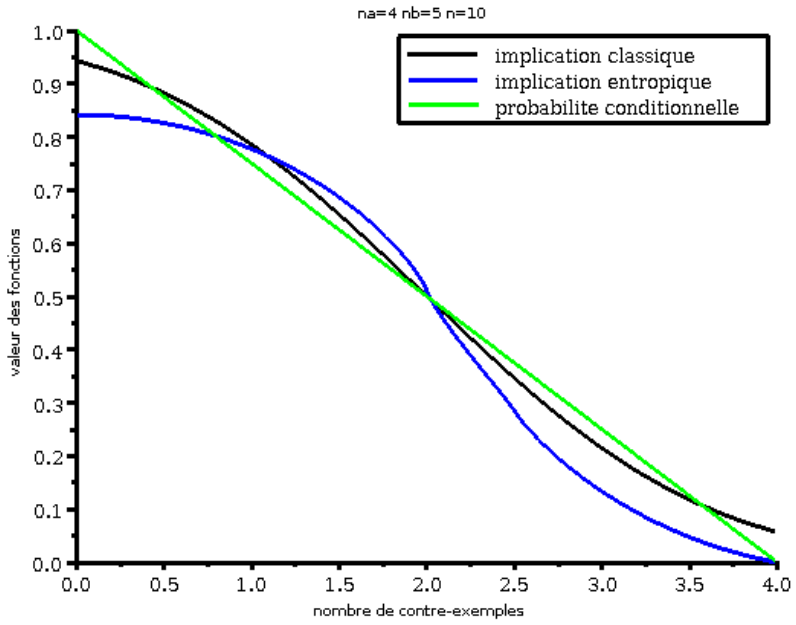


f) Propriété P_6 : accepter la variété de types de variables

La mesure semble devoir être suffisamment souple et analytiquement générale pour qu'elle soit opératoire sur différents types de variables, dont des variables floues, et pour que son calcul supporte le mélange de variables de natures différentes. Par exemple, si un indice est défini pour le cas non binaire, sa restriction au cas binaire doit coïncider avec l'indice, fournissant la mesure, défini dans ce dernier cas. A notre connaissance, tous les indices d'implication, hors de l'A.S.I., ne sont opératoires que sur les variables binaires.

Exemple : L'indice l'implication a été étendu à des variables non binaires tout en supportant sa restriction au cas binaire. C'est le cas des variables numériques, modales et intervalles. On pourra se reporter à (LAGRANGE 1998) ou à (GRAS et al. 2009). Ces extensions ouvrent un champ d'applications d'autant plus large que le logiciel CHIC accepte l'association de types de variables différents au cours du même traitement des données.

Figura 5. Comparaison des variations de l'implication classique, l'implication entropique et la probabilité conditionnelle en fonction du nombre de contre-exemples



g) Propriété P_7 : assurer une confiance statistique discriminante

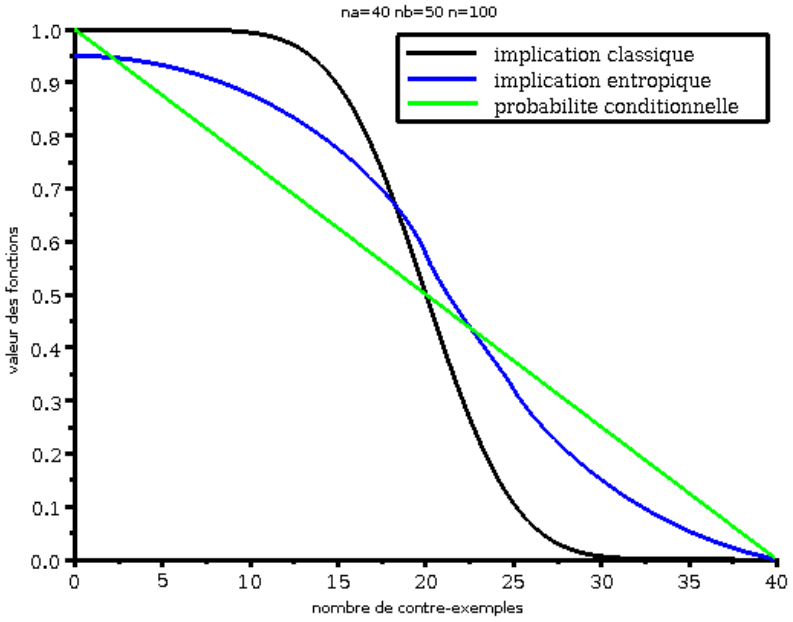
L'exemple cité dans l'introduction, où le cardinal de la population croît d'une situation (n=10) à la suivante (n=1000), montre bien que la mesure de qualité devrait être sensible à ce cardinal en étant explicitement fonction de lui. Puisque la fonction de l'analyse de données est de nature inductive pour esquisser une convergence de la contingence vers la nécessité, il s'agit d'assurer une **confiance statistique** à l'élection des règles, tout en respectant une certaine **qualité discriminante** de l'indice de base. (LERMAN, AZE, 2004), sur la base de nos travaux A.S.I., obtient un indice discriminant en utilisant, à notre avis, un artifice de centrage-réduction de l'indice de base, ce qui ne fait qu'écarter de 1 les valeurs numériques s'y écrasant.

Exemple 1 : cf. les figures 5, 6 et 7 pour apprécier la différence, montrée en ordonnée dans un graphe, entre la probabilité conditionnelle, l'intensité d'implication classique et l'intensité entropique. On fait varier en abscisse le nombre de contre-exemples et avec le même rapport d'homothétie (x1, x10, x1000) les différents paramètres : n_a, n_b et n. On constate dans ces

figures que l'intensité d'implication classique varie avec n alors que la confiance et l'indice d'inclusion entropique, élément de l'intensité *d'implication-inclusion*, sont invariantes avec *n*

$i(a,b) = \left(\left[1-h_1(t)\right]\left[1-h_2(t)\right]\right)^{\frac{1}{2}}$ où $h_1(t)$ et $h_2(t)$ sont respectivement les entropies conditionnelles de *b* sachant *a* et de *non a* sachant *non b*.

Figura 6. Comparaison des variations de l'implication classique, l'implication entropique et la probabilité conditionnelle en fonction du nombre de contre-exemples



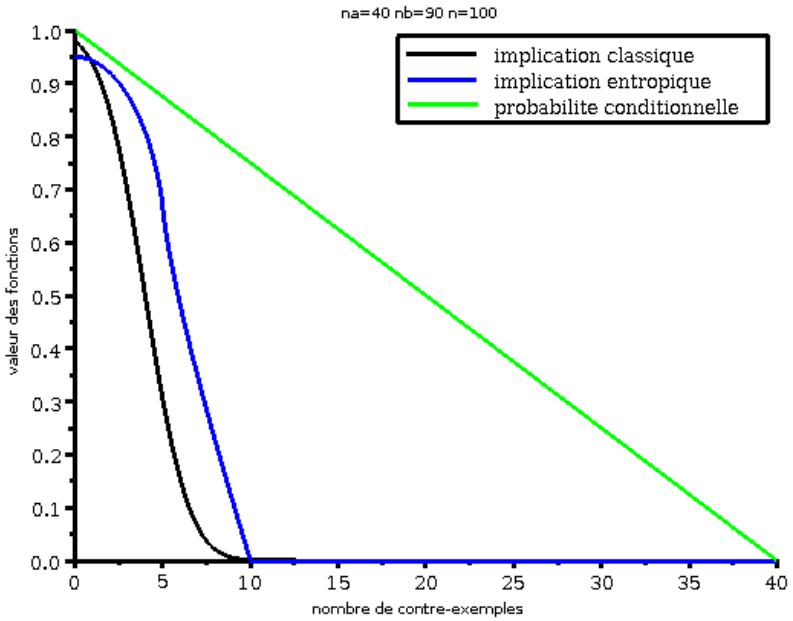
Le coefficient $1-\frac{1}{2\sqrt{n}}$. croissant avec n est indicateur de confiance qui ne peut que renforcer le critère de qualité de la règle a=>b dont rend compte l'indice d'inclusion. Plus il est grand, plus fiable est le rôle de l'indice entropique pour signifier la qualité de la règle. Aussi, nous avons adopté,comme indice d'implication entropique, le nombre : $\Psi(a,b) = (1-\frac{1}{2\sqrt{n}})i(a,b)$.

Par contre, comme le montrent les figures 8 et 9, discriminante pour des valeurs de *n* de l'ordre de quelques centaines, l'intensité d'implication classique $\varphi(a,\bar{b})$, contrairement à l'implication entropique, ne l'est plus lorsque *n* est très grand (rapport d'homothétie : x1 puis x100). En conséquence, cette intensité est voisine de 1 le long d'une plage importante de contre-

exemples, propriété que conserve, uniformément, la fréquence conditionnelle. Ce qui est *préjudiciable* à l'attente de ce que nous avons appelé **« confiance statistique ».** Ce problème de discrimination est aussi évoqué et traité à travers l'article de J. Blanchard et als (BLANCHARD et al., 2005b) et dans (GRAS et al. 2001a) et, rappelons-le, résolu avec l'implication entropique.

Nous avons rencontré un problème comparable d'inadéquation d'un modèle lorsque le nombre des données croît au-delà des situations habituelles : il s'agit du test du χ^2 d'ajustement d'une loi empirique à une loi théorique.

Figura 7. Comparaison des variations de l'implication classique, l'implication entropique et la probabilité conditionnelle en fonction du nombre de contre-exemples.

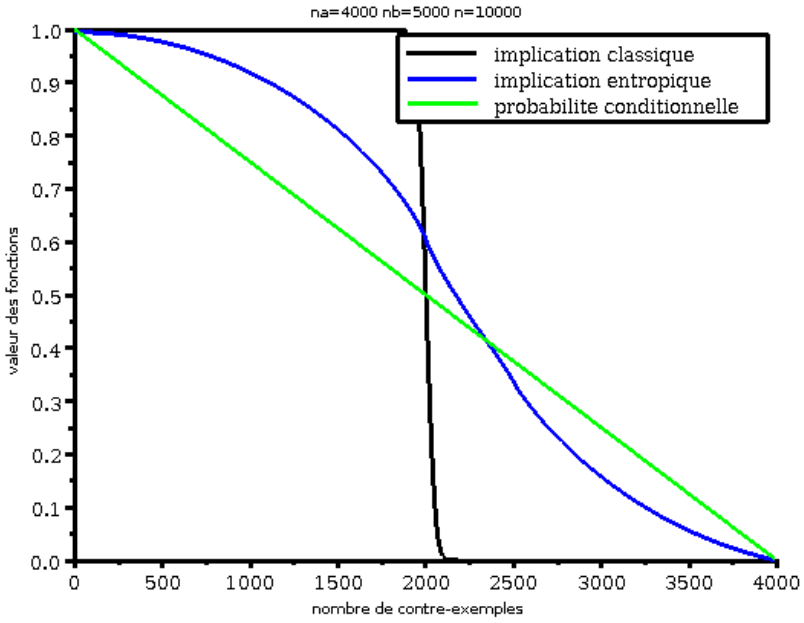


Exemple 2 L'indice appelé « Multiplicateur de cote » (LALLICH S. et al, 2004) : $m = \frac{n_a - n_{a\wedge\bar{b}}}{n_b . n_{a\wedge\bar{b}}} n_{\bar{b}}$ présente de bonnes qualités que nous lui reconnaissons et qui sont exprimées dans le texte des auteurs mais il conduit, comme la confiance, au même inconvénient de rester

invariant par dilatation des données. De plus, il est indépendant de n donc ne varie pas avec l'effectif de la population, bien que, non explicitement, n intervienne comme majorant de chacun des paramètres en présence dans l'indice m.

Par exemple, dans le cas des situations représentées par les figures 7 et 9, si $n_{a\wedge\bar{b}} = n/50$, m=19/9 et si $n_{a\wedge\bar{b}} = n/10$, m = 1/3 quelles que soient les situations.

Figura 8. Comparaison des variations de l'implication classique, l'implication entropique et la probabilité conditionnelle en fonction du nombre de contre-exemples

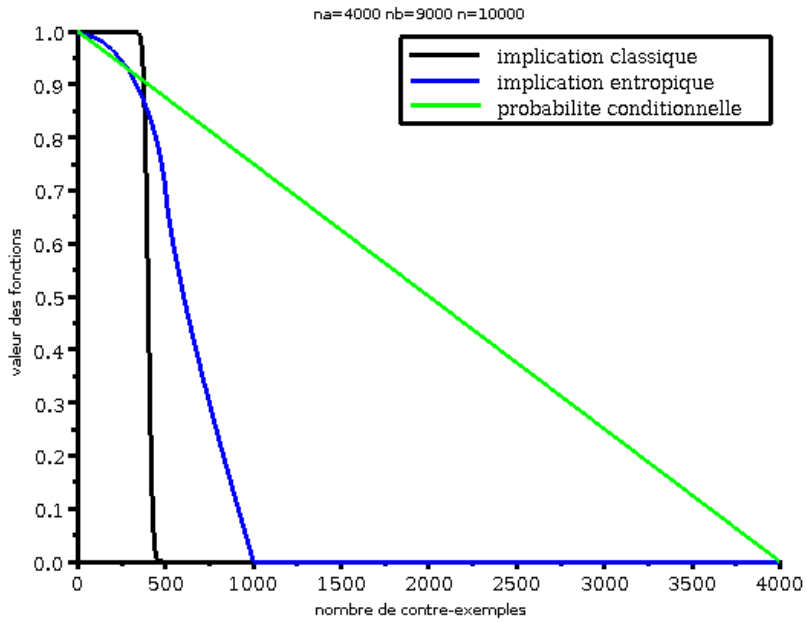


h) Propriété P_8 : permettre l'adéquation de la mesure avec la contraposée de la règle

Afin de renforcer le caractère quasi-implicatif de *a* vers *b* i.e. la quasi-inclusion de l'ensemble des exemples de *a* dans celui des exemples de *b*, **la mesure de *a*⇒*b* devrait être couplée à celle de sa contraposée *non b*⇒*non a***, leurs valeurs associées pouvant être très différentes. Nous soulignons l'intérêt de cette propriété pour assurer à la règle une propriété prédictive car significative d'un fondement causal. En effet, si l'implication directe ***a*⇒*b*** relie la cause ou les causes conjointes à l'effet, la contraposée ***non b*⇒*non a*** nous indique que la disparition

de l'effet s'accompagne de l'extinction de la ou des causes, ce qui est manifestement et complémentairement informatif et sémantiquement satisfaisant dans une visée causale.

Figura 9. Comparaison des variations de l'implication classique, l'implication entropique et la probabilité conditionnelle en fonction du nombre de contre-exemples.



Exemple 1 : Les indices issus directement de l'A.V.L. (Algorithme de la Vraisemblance du Lien de I.-C. Lerman) confèrent, autant pour les règles de ressemblance que pour celles d'implication, les mêmes valeurs respectives aux similarités des variables *a* et *b* qu'à leur négation, et à une implication qu'à sa contraposée. C'est donc le cas de l'indice d'implication

et par conséquent de l'intensité d'implication classique. Avec regret, l'information tirée de l'implication et de sa contraposée est la même.

Exemple 2 : Par contre, la confiance et, comme on le sait, l'indice entropique défini en A.S.I. et, a fortiori l'intensité d'implication-inclusion (cf. (c)) en évaluant différemment les deux règles, prennent en compte la contraposée. On examine dans d'autres articles plus longuement l'intérêt de cette propriété par rapport à l'objectif de recherche de causalité entre les variables. Il est bien évident que les qualités respectives d'une implication et de sa contraposée confèrent une garantie plus forte d'une relation causale entre a et b.

i) Propriété P_9 : définir des algorithmes simples et intelligibles

La formule et les algorithmes conduisant à la mesure de la règle doivent présenter un triple caractère **:**

- d'une part, leur **complexité** ne doit pas conduire à des temps d'exécution trop longs, en particulier lorsque le fichier de données comporte un grand nombre de variables,
- d'autre part, leur définition doit pouvoir être **appréhendée intuitivement** si l'on souhaite que les valeurs obtenues gardent une **signification** permettant une interprétation aisée et partagée ; Lenca P. (LENCA et al., 2004) parle à ce sujet d**'intelligibilité,**
- enfin, les algorithmes doivent être programmables de telle façon que l'utilisateur ait accès aux résultats, puisse agir sur les seuils et se donne ainsi le moyen de pratiquer les analyses lui-même (cf. le logiciel CHIC pour le traitement de l'A.S.I. (COUTURIER 2008, COUTURIER ; AG ALMOULOUD, 2009 ; et RATSIMBA-RAJOHN, 2009)).

Par exemple, la fréquence conditionnelle, un indice de type Loevinger ou ses dérivés possèdent les deux premiers caractères. En revanche, la J-mesure, la mesure Zhang, l'intensité d'implication et, a fortiori l'implication-inclusion (cf. paragraphe (g)), s'ils ne conduisent pas à des calculs complexes, ont perdu dans leur formulation le caractère intuitif qui a guidé pourtant leurs définitions.

De plus, lorsque l'on considère les conjonctions de variables en prémisses, le nombre de règles croît, par exemple, en n^2 lorsque sont évaluées les règles à deux prémisses. L'indice d'Agrawal (AGRAWAL 1993), sur la base de la confiance, conduit ainsi à un nombre inextricable de règles, en dépit des seuils fixés. L'idée qu'expriment (COUTURIER et al, 2004) en introduisant un indice dit d'originalité permet de réduire très sensiblement cet écueil car cet indice fait le tri entre les conjonctions logiques envisageables pour ne retenir que celles qui conduisent à l'amélioration d'un indicateur de qualité.

j) Propriété P_{10} : enrichir l'accès à la connaissance par représentation graphique des règles et des métarègles

Les propriétés analytiques de la mesure devraient permettre l'émergence et l'interprétation de propriétés d'un « tout », à travers une structure obtenue sur l'espace des règles ou des métarègles, un « tout » qui ne soit pas la somme des propriétés des parties que sont les règles (cf. note 9 du (c)). Un « tout » constitué de la concomitance asymétrique et emboîtée des règles produites, accèdant à ce que J.Piaget appelle un niveau supérieur de complexité comme le serait une composition de fonctions. Ce sera par exemple, l'apparition (l'émergence) d'un modèle de l'élève qui se dégagerait d'une structuration de ses comportements dans des situations-problèmes (cf. GRAS, 1979, LAHANIER-REUTER, 1999 et CROSET, 2007). Mais cette émergence nécessite une compatibilité :

- d'une part, entre la structure d'ordre ou de préordre partiels induite sur les variables par la donnée des règles implicatives et celle de représentations graphiques des relations les associant,
- d'autre part, comme nous le verrons dans (k), avec les structures induites par les règles éligibles sur l'ensemble des variables actives ou principales et l'ensemble des variables exogènes ou supplémentaires.

Ces représentations se font, dans le cas de l'intensité d'implication selon un ***graphe*** orienté, pondéré et sans cycle, dans un espace affine (2 dimensions) : lisibilité et fidélité sont donc requises. Par exemple, la représentation graphique par le logiciel CHIC de l'ensemble des règles implicatives conduit à laisser isolées des variables non liées à d'autres au-dessus d'un seuil de qualité donné par l'utilisateur(0,5 en l'occurrence dans CHIC par défaut). En cela, l'A.S.I. se distingue avantageusement des recherches qui se limitent à l'étude des relations deux à deux des variables ou de leurs conjonctions à l'aide d'un indice différent d'une recherche à une autre, c'est-à-dire essentiellement de l'étude de « parties ». A notre connaissance seules les représentations graphiques :

- **les treillis de Galois** pour des données booléennes se construit sur le principe d'inclusion logique, dans un modèle algébrique non statistique, de variables ou de sujets ; mais les liens entre groupes de variables peuvent exister alors qu'aucun sujet ne les vérifie, l'essentiel étant qu'aucun ne les contredise ;

- **les arbres de décision ou graphes d'induction**, structurent le croisement de variables qualitatives à prédire à l'aide d'un espace de prédicteurs sur un ensemble de sujets. Pour cela, on choisit à chaque nœud de l'arbre les meilleures partitions parmi les attributs d'une variable. Le but est d'identifier un attribut à prédire (RITSCHARD et al., 2007 et 2009)

- **et les réseaux bayesiens**, d'origine anglo-saxonne (i.e. PEARL, 1988), fonctionnent dans un modèle fréquentiel au moyen de l'indice « probabilité conditionnelle ». Ils présentent une propriété organisatrice de l'ensemble des variables principales (voir une approche synthétique, comparative avec l'ASI et critique dans (CADOT, 2009)). Ils structurent en un graphe des variables booléennes dont les arcs représentent des liens probabilistes entre des faits vérifiés par ces variables (ex : « Filippo a pris un café ce matin »). Les relations causales recherchées de type « si *a* alors *b* », « si *a* alors non *b* », « si non *a* alors *b* », « si non *a* alors non *b* », sont quantifiées par des probabilités a priori de « causes » et les probabilités conditionnelles vers d'éventuelles « conséquences » qui en découleraient. La formule de Bayes est utilisée pour cette

réactualisation des probabilités de « causes », par exemple au cours d'observations répétées. D'où la complexité des calculs récursifs sur tout ou partie du graphe, l'impossibilité de donner du sens à des chemins du graphe (le « tout » reste ici un ensemble de « parties », la structure est locale), l'indépendance par rapport à n. L'observation d'une cause ou de plusieurs causes n'entraîne pas systématiquement l'effet ou les effets induits, mais modifie seulement la probabilité de les observer. Leur valeur prédictive est efficace mais plus en économie-gestion qu'en sciences humaines.

A l'opposé, le ***graphe implicatif*** (GRAS, 1979) ou la ***hiérarchie cohésitive*** (GRAS et al, 2002 et 2003) conduisent à l'étude de structure, réorganisant ces « parties » en un « tout » cohérent au sein de modèles probabilistes. Les méthodes classificatoires ou factorielles symétriques le permettent mais l'ASI, méthode non symétrique, dispose du privilège d'ouvrir la voie dans cet autre cadre. En effet, cette absence de symétrie confère aux structures obtenues une qualité dynamique. Comme le dit J.L. Deneubourg dans (BENKIRANe, 2002, p. 107), au sujet d'auto-organisation des termites : « C'est la dynamique du jeu entre certaines règles –par exemple, pour les termites, « déposer » ou « cela sent »- qui produit la structure générale ».

Relativement au graphe implicatif, le problème de la non-transitivité d'une suite de règles doit y être pris en compte. Ainsi l'ordre des arcs d'un chemin d'un graphe implicatif (ex. : $a \rightarrow b \rightarrow c\ldots$) doit être compatible avec la suite des règles admissibles $a \Rightarrow b$, $b \Rightarrow c$ *et* $a \Rightarrow c,\ldots)$; ainsi, dans le même exemple, si l'utilisateur exige une transitivité à un seuil donné, certaines liaisons apparues sur le graphe vont disparaître de celui-ci. Cependant, CHIC peut tracer les arcs transitifs qui subsistent.

De plus, en ce qui concerne la hiérarchie cohésitive, la propriété classificatoire définie en A.S.I. – ***la cohésion*** - pour hiérarchiser les métarègles présente un avantage indéniable par rapport aux critères choisis généralement en classification. En effet, alors que ces critères conduisent à certains artefacts comme *rassembler dans la même classe des éléments de faible liaison, voire*

s'opposant, la cohésion évite cet écueil en rejetant la formation d'une classe si la mesure de la liaison implicative n'indique plus une certaine dépendance.

Afin de valider la qualité des structures obtenues au moyen de ces deux représentations, des indices de qualité ont été définis : la variance implicative pour les graphes implicatifs (GRAS et al., 2009) et la significativité des niveaux de la hiérarchie pour la hiérarchie cohésitive (GRAS et al, 2004). Nous renvoyons le lecteur à ces différents travaux.

Nous renvoyons, de même, le lecteur à la comparaison établie par (RITSCHARD et al., 2007) entre l'arbre d'induction selon le critère de l'entropie décentrée et l'indice d'implication. Il apparaît que l'entropie décentrée favoriserait le développement de l'arbre implicatif alors que l'indice d'implication serait préférable pour l'attribution des conclusions aux feuilles des règles constituant les branches de cet arbre.

k) Propriété P_{11} : tendre vers une structuration duale entre variables actives et variables supplémentaires

Il nous semble important et enrichissant, comme en analyse factorielle, d'associer la structure induite sur l'ensemble des variables actives ou principales (i.e. participant à la recherche des règles d'association) par l'élection de règles et celle induite sur l'ensemble des variables supplémentaires (des sujets, des groupes de sujets ou des descripteurs par exemple) relativement à chaque règle. La recherche de contribution d'une variable supplémentaire à l'existence ou l'émergence d'une règle doit pouvoir être calculée sur la base de sa satisfaction ou sa non-satisfaction à la règle. L'ordre entre les contributions à une règle donnée définit un préordre total spécifique sur l'ensemble des variables supplémentaires, allant de la variable contribuant le plus à celle dont la contribution est minimale. La possibilité est alors donnée d'extraire les **individus limites** (border subjects) intervenant de façon exceptionnelle dans une relation contrairement à l'ensemble des autres sujets. La connaissance de leurs descripteurs pourra contribuer à la donnée de sens à un chemin ou une classe de variables et d'identifier ou de désigner leur rôle dans leur constitution. A notre connaissance, seule l'A.S.I. permet l'étude simultanée des structures induites par les statuts respectifs « principales-supplémentaires » de l'une sur l'autre. Mieux, grâce à cette dualité originale, comparable à la dualité en analyse

factorielle, entre l'espace des variables et celui des sujets, elle conduit à une topologie métrique induite sur ce dernier (GRAS et al, 2009).

Regard croisé avec une analyse de critères comparable

Nous ne hiérarchisons pas ces onze propriétés, bien qu'elles n'aient pas toutes la même importance à nos yeux. A fortiori, nous ne les dotons pas de pondérations qui viseraient la mise en évidence d'une mesure idéale et optimale, trop de points de vue pouvant interférer, se conjuguer ou s'opposer. En réalité, c'est l'utilisateur qui, en fonction de ses objectifs et de la sémantique qu'il privilégie, est le meilleur juge : tantôt telle propriété sera primordiale, tantôt elle sera plus secondaire. Nous donnons ici simplement notre propre réponse quant à une mesure de qualité, en tentant de montrer et de justifier notre choix, eu égard aux propriétés ci-dessus. Nous le faisons en termes d'indices certes, mais aussi en montrant leur élargissement à la fois sur le plan de la variété des types de variables traités, sur celui de règles généralisées (règles de règles), sur celui de l'importance accordée aux structures et sur celui de la considération des sujets dans leur responsabilité à l'élection de règles simples ou généralisées. Il est bien évident que nos propres choix ont tendu et tendront encore à satisfaire les propriétés énoncées plus haut, ce qui peut paraître restrictif et peu critique à l'égard de ces choix. Mais une attitude contraire se traduirait par de l'incohérence !

Examinons maintenant une étude intéressante portant sur des critères de choix d'indices implicatifs de la littérature de fouille des données, y compris l'A.S.I.. P. Lenca dans (LENCA et al.,2004), après avoir rappelé quelques indices d'association implicative (dont l'intensité d'implication classique et l'intensité entropique, les deux versions actuelles de l'A.S.I.) retient 8 critères d'évaluation :

g_1 : traitement non symétrique de a et b ;

g_2 ; décroissance avec n_b ;

g_3 : évaluation des situations de références, indépendance : la mesure doit être fixe (0 ou 1) dans le cas d'indépendance entre *a* et *b* ;

g_4 : évaluation des situations de références, règles logique : la mesure doit être maximum ou tout au moins très forte en cas de règles logiques (aucun contre-exemple) ;

g_5 : la mesure est non linéaire en $n_{a\wedge\bar{b}}$ au voisinage de 0 ;

g_6 : prise en compte de *n* :

g_7 : facilité à fixer un seuil d'acceptation de la règle ;

g_8 : intelligibilité de la mesure.

Ces différents critères répondent assez fidèlement aux propriétés que nous venons d'examiner. Il s'agit de g_4 et g_7 apparentés à **P_1**, g_5 à **P_2**, g_2 à **P_3** et **P_5,** g_6 à **P_7**, g_8 à **P_9**, g_3 à **P_2**, g_1 à **P_4**. Mais on constate que ne sont pas envisagées les propriétés, satisfaites ou poursuivies plus ou moins en A.S.I. :

- **P_6** où sont évoqués les autres types de variables ;
- **P_8** où la contraposée est prise en compte, condition favorisant la fonction causale de la règle ;
- **P_{10}** où la nécessité de créer des modèles graphique facilitant la lecture et l'analyse globale des données en terme de structures de règles et/ou métarègles ;
- **P_{11}** où le traitement dual des variables supplémentaires est occulté.

Nonobstant ces écarts avec nos propriétés, une étude intéressante, à partir d'une distance définie sur l'ensemble des 20 indices retenus par P. Lenca, permet de faire apparaître les proximités entre ces indices en fonction des intérêts que privilégierait l'utilisateur. On note en particulier que l'indice Zhang est proche des deux indices à la base des intensités d'implication classique et entropique. Et que celles-ci sont remarquablement placées si l'utilisateur accepte l'apparition lente d'un certain nombre de contre-exemples et la non-linéarité des indices. Par contre, elles le sont moins si l'utilisateur exige une chute brutale de l'indice dès l'apparition d'un ou de quelques contre-exemples indépendamment des exemples en jeu.

Conclusion

Nous avons établi une liste de propriétés susceptibles de définir la qualité de règles d'association, voire la qualité de la mesure à attacher à ces règles. Elles sont censées répondre à des attentes sémantiques **qui, de ce fait, y puisent leur signification,** justifiant également les choix épistémologiques, certaines propriétés étant relatives à la ressemblance entre variables, et d'autres, de façon plus spécifique, à des règles implicatives. Nous avons puisé, sans exhaustivité, dans la littérature des mesures de qualité de règles d'association, celles qui paraissaient satisfaire pour partie ces propriétés, celles qui semblaient ne pas les satisfaire. Des illustrations numériques et graphiques ont cherché à appuyer nos argumentations. A la faveur de cette revue d'indices de qualité, nous nous apercevons que peu de mesures satisfont l'ensemble des propriétés, y compris l'A.S.I., ne serait-ce que parce que certaines vont s'opposer à d'autres ou tout au moins parce que renforcer telle propriété conduit à affaiblir telle autre. Généralement, les mesures de qualité satisfont l'attente commune et légitime d'être sensibles à certains cardinaux en jeu. Certaines (LALLICH et al, 2007) permettent une analyse statistique fine de l'indice retenu dans le modèle implicatif. Mais l'A.S.I. présente une spécificité, moins commune, d'une part, de mettre en évidence ce qui est *surprenant, non nécessairement attendu*, des « pépites de connaissance » comme dit Y.Kodratoff (KODRATOFF, 2000)[15], et, d'autre part, de refuser les règles triviales.

Nous avons parallèlement rappelé les mesures que nous choisissons dans nos groupes de recherche, l'intensité d'implication statistique classique ou l'intensité entropique. Nous avons essayé d'objectiver[16] nos choix de concepts et d'algorithmes tout en montrant leur relation

[15] Nous rejoignons en cela René Thom (Thom R., 1980) : « …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 ».

[16] comme nous le faisons en didactique des mathématiques par une recherche optimale de donnée de sens : pourquoi ce concept ? à quelles problématiques répond-il ? quels problèmes résout-il ?

positive, neutre ou négative avec les propriétés retenues. Par exemple, la conception de l'intensité entropique est liée à la reconnaissance critique de la faiblesse de transparence de l'intensité classique dans bon nombre de situations. En même temps, s'appuyant sur la contraposée et la qualité de l'information, l'intensité entropique se présente mieux affûtée pour aborder les causalités sous-jacentes ; certes, causalité singulière car extraite de la contingence mais susceptible d'explicabilité sinon de prédictibilité. Cet écart épistémologique entre les deux modèles de mesure nous semble cohérent avec la recommandation de J.-M. Leblond (LEBLOND, 1981, 77) : « Maîtriser un concept scientifique, c'est d'abord connaître les limites de sa validité et de sa pertinence ». *Validité*, par rapport à des critères objectifs ; *pertinence* par rapport à la sémantique originelle ; mais aussi, *vigilance* épistémologique critique vis-à-vis des créations ad-hoc non justifiées sur les plans sémantique et épistémologique.

L'ASI, nous l'avons vu à travers la propriété $\mathbf{P_{10}}$, offre aussi l'originalité de dépasser la simple mesure de qualité de règles par l'étude des structures qu'elles peuvent constituer, prises dans leur ensemble, y compris celle qui exprime une dualité topologique entre les ensembles de variables et de sujets. Notre modélisation d'une implication statistique structurante, comme le dit René Thom (THOM, 1980), « est un pari. Il y a une mise et un gain : la mise c'est la justification *a priori*, le gain celle *a posteriori* ». Nous avons construit cette mise, pièces par pièces, par les propriétés. Le gain apparaît dans les multiples applications multidisciplinaires de l'A.S.I.. Il est indéniable, comme le montrent les textes qui accompagnent celui-ci, dans (GRAS et al, 2008) et (GRAS et al, 2009).

Nous avons évoqué la richesse du passage de l'étude des « parties » à celle d'un « tout ». Cette étude opère dialectiquement au sens où l'interaction critique y est constamment présente. Les significations éparses, a priori hétérogènes, se recomposent en une signification globale, de type « organique », comme il a été fait pour recomposer en un modèle d'élève ses comportements au cours d'une situation-problème (CROSET, 2008) ou en un modèle d'enseignant sur la base de ses choix épistémologiques (BAILLEUL, 1994), (BODIN ET GRAS, 1999), (GRAS et al, 2009) ou des conceptions du hasard (LAHANIER-REUTER, 1999). Comme le dit, L.Sève, (SEVE et al, 2005), souvent cité ici, « …tel celui de la cause et de l'effet, tout et partie ne forment qu'un *seul et même concept* : celui du *rapport tout/partie* »

(p.129) et plus loin : « …le tout n'est égal aux parties qu'en tant qu'elles sont *« prises ensemble.*. » (p.130).

De ce fait, nous sommes alors convaincus que la dialectique « quantité-qualité » conduit à la faveur de seuils quantitativement franchis (population, variables, indice) à des *sauts qualitatifs* dans la compréhension du « tout », comparables au passage de l'eau à la glace par abaissement de la température. C'est l'organisation des parties en système, les rapports entre elles qui forment la source explicative de l'émergence de la nouvelle compréhension du « tout ». Ceci tempère, voire oblitère « l'espoir réductionniste de trouver dans les propriétés des constituants de quoi expliquer celles du tout » (ATLAN,1986). On retrouve, semble-t-il, ce même rejet d'un *Tout* réduit à ses parties dans le dialogue sur l*'Etre* conduit par Platon dans *Parménide*. C'est la connexion de l'ensemble des parties qui conduit à l'émergence d'une propriété du tout à un certain niveau d'organisation. Etant entendu qu'une extension quantitative peut conduire, à un niveau supérieur, par une nouvelle organisation, à une nouvelle intégration au cœur d'une propriété peut-être plus générale. Toute analyse d'un chercheur travaillant à l'aide de l'A.S.I. et de CHIC sur un corpus de données doit toujours viser et atteindre ce type de synthèse, enrichie en outre par l'apparition de certaines propriétés inattendues. Il doit impérativement fonctionner dialectiquement, dans son analyse, en enrichissant le sens du tout, c'est-à-dire du complexe, par celui de la partie, c'est-à-dire du simple, et réciproquement, en réduisant les contradictions inhérentes à cette logique. Ainsi l'analyste passera de l'interprétation d'arcs juxtaposés, à celle du chemin les portant, de celle de chemins conjoints à celle d'une grappe de chemins et vice versa dialectiquement et synthétiquement. Idem pour les classes de la hiérarchie cohésitive. Edgar Morin dirait que la pensée (ordinaire) se doit aussi de séparer et opposer les phénomènes au lieu de les rapprocher et les envisager dans leur complexité.

C'est tout l'intérêt qu'apportent le graphe implicatif et la hiérarchie cohésitive. Ces deux modes de représentation, comme l'est une phrase pour exprimer dynamiquement une idée, un sentiment, permettent de « dessiner une topologie du sens » (GAUDIN, 05), ne supposant ni additivité, ni proportionnalité entre le sens de leurs composants et celui de leur tout. En résumé,

l'A.S.I. nous semble pouvoir conférer à un ensemble de données une structure dynamique, animée par sa dissymétrie de traitement, une structure systémique et extensive. C'est en cela qu'elle puise toute sa fécondité.

On voit alors que la problématique de la conception d'une mesure de qualité de règle, même satisfaisant au mieux les propriétés énoncées ici, n'est qu'un fragment de l'élaboration d'une méthode de recherche et d'analyse de règles d'association non symétrique. Quelquefois, certains chipotages et chicaneries sur leur choix sont secondaires. L'A.S.I. a pour objectif de dépasser le stade de la seule élection de règles de bonne qualité. Elle se singularise par rapport aux autres choix élus par la **double structuration des règles, par leurs représentations et par la dualité variables-sujets.**

Nos horizons se tournent vers de nouveaux développements de l'intensité d'implication afin d'en étudier la stabilité et pour qu'elle puisse répondre à de nouveaux types de variables. D'autres problèmes soulevés par les applications, conduisent en effet à de nouvelles pistes de recherche : par exemple, d'une part, les variables rangs (analyse des préférences), les variables continues, les données floues, la dualité topologique évoquée ci-dessus et, d'autre part, le problème que posent les tableaux aux données manquantes, ou encore celui qui se présente lorsque le nombre de variables devient excessif au point de complexifier l'interprétation des représentations. Il ne fait pas de doute que nous chercherons toujours à respecter les propriétés que nous aurons jugées fondamentales et sémantiquement justifiées.

Références

Agrawal R. et al. (1993), Mining association rules between sets of items in large databases, *Proc. of the ACM SIGMOD'93.*

Ag Amouloud S., (1992), *L'ordinateur, outil d'aide à l'apprentissage de la démonstration et de traitement de données didactiques*, Thèse de doctorat de l'Université de Rennes 1.

Atlan H. (1986), *A tort ou à raison,* Seuil.

Aze J. et Kodratoff Y. (2001), Evaluation de la résistance au bruit de quelques mesures d'extraction de règles d'association , *Extraction des connaissances et apprentissage,* Hermès, Vol 1, n° 4, 2001, p. 143-154

Bailleul M. (1994), *Analyse statistique implicative : variables modales et contributions des sujets. Application à la modélisation de l'enseignant dans le système didactique*, thèse, Université de Rennes I.

Benkirane R. (2002), *La complexité, vertiges et promesses, Entretiens avec E. Morin, I.Prigogine, F. Varela,…* Le Pommier

Bernard J.-M., Poitrenaud S. (1999), L'analyse implicative bayesienne d'un questionnaire binaire : quasi-implications et treillis de Galois simplifié, *Mathématiques, Informatique et Sciences Humaines* 147, p.25-46.

Blanchard J., Guillet F., Briand H. (2002), L'intensité d'implication entropique pour la recherche de règles de prédiction intéressantes dans des séquences de pannes d'ascenseur, *Extraction et gestion des connaissances*, Hermès, p. 77-88.

Blanchard J., Guillet F., Briand H. et Gras R., (2005a), IPEE : Indice Probabiliste d'Ecart à l'Equilibre pour l'évaluation de la Qualité des Règles, *RNTI-E-5, Cépaduès Editions*, p. 391-395.

Blanchard J., Kuntz P., Guillet F. et Gras R. (2005b), Mesure de la qualité des règles d'association par l'intensité d'implication entropique, *Mesure de Qualité pour la Fouille de Données, RNTI-E-1, dir. H.Briand, M.Sebag, R.Gras, F.Guillet, Cépaduès Editions*., p. 33-44.

Bodin A., Gras R. [1999], Analyse du préquestionnaire enseignants avant EVAPM-Terminales, *Bulletin n°425 de l'Association des Professeurs de Mathématiques de l'Enseignement Public*, p. 772-786, Paris

Brin S ., Motwani R., Silverstein C. [1997], Behond market baskets : generaling association rules to correlations, *in ACM SICMOD / PODS 97Joint conference*, p. 265-276.

Cadot M., (2009), Graphe de règles d'implication statistique pour le raisonnement courant. Comparaison avec les réseaux bayesiens et les treillis de Galois, *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*, p.223-250.

Couturier R, Gras R., et Guillet F. (2004), Reducing the number of variables using implicative analysis In International Federation of Classification Societies, IFCS 2004, Springer Verlag: Classification, Clustering, and Data Mining Applications, ISBN 3-540-22014-3, Chicago, USA, July 2004, p. 277--285.

Couturier, R. (2008). Statistical implicative analysis. In CHIC : Cohesive Hierarchical Implicative Classification, *Volume 127 of Studies in Computational Intelligence, Springer Verlag, ,* p. 41–5.

Couturier R. et Ag Almouloud S. (2009), Historique et fonctionnalités de CHIC, *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 p.279-293

Croset M.-C.,(2007), Un modèle d'élève par l'analyse statistique implicative. Prise en compte du contexte algébrique, *Nouveaux Apports Théoriques à l'Analyse Statistique Implicative et Applications, 4th International Meeting on Statistical Implicative Analysis*, *Ed. R.Gras, P.Orus, B.Pinaud, P.Gregori*, p.211-220

Fleury L. (1996), *Extraction de connaissances dans une base de données pour la gestion des ressources humaines,* Thèse d'Université, Nantes, 1996

Ganascia J.G. (1991), CHARADE : Apprentissage de bases de connaissances, *Induction symbolique-numérique à partir de données,* Cépaduès Éditions, 1991

Gaudin P., (2005), Y a-t-il de la non linéarité en sémantique ?, dans *Émergence, complexité et dialectique*, Odile Jacob, Paris, p. 279-288

Goodman R.M. et Smyth P. (1989), The induction of probabilistic rule set. The ITRULE algorithm, *Proceedings of sixth international conference on machine learning*, 1989, p. 129-132

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'État, Université de Rennes 1, 1979.

Gras R. et Larher A (1993), L'implication statistique, une nouvelle méthode d'analyse de données, *Revue Mathématiques, Informatique et Sciences Humaines n°120*, p.5-31.

Gras R. et al. (1996). *L'implication statistique. Nouvelle méthode exploratoire de données.* Grenoble, La Pensée Sauvage.

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

Gras R., Kuntz P., Couturier R. Guillet F. (2001a), Une version entropique de l'intensité d'implication pour les corpus volumineux, *Extraction et Gestion des Connaissances,* Vol. 1, Hermès, p. 69-80.

Gras R., Kuntz P., Briand H. (2001b), Les fondements de l'analyse statistique implicative et leurs prolongements pour la fouille de données, *Mathématiques et Sciences Humaines* n° 154-155, p. 9-29.

Gras R., Diday E., Kuntz P, Couturier R. (2001c), Variables sur intervalles et variables-intervalles en analyse statistique implicative, *Actes VIIIèmes Rencontres de la S.F.C.*, Université de Pointe-à-Pitre, 2001.

Gras R., Kuntz P., Briand H., Couturier R. (2002), Hiérarchie de règles généralisées et notion de variable supplémentaire en analyse statistique implicative, *Actes des IXèmes Rencontres de la Société Francophone de Classification,* Université de Toulouse, 2002, p. 211-214.

Gras R., Kuntz P., Briand H. (2003), Hiérarchie orientée de règles généralisées en analyse implicative, *Extraction et Gestion des Connaissances 2003*, Vol 1, Hermès, p. 145-158.

Gras R., Couturier R.., Blanchard J., Briand H., Kuntz P., Peter P., (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,* p 3-32

Gras R, Kuntz P.et Régnier J.C. (2004), Significativité des niveaux d'une hiérarchie orientée, *Classification et fouille de données, RNTI-C-1, Cépaduès- Editions, ISBN 2.85428.667.7,* p 39-50.

Gras R. et Kuntz P. (2005), Discovering R-rules with a directed hierarchy, *Soft Computing, A Fusion of Foundations, Methodologies and Applications,Volume 1, ISSN 1432-7643,* Springer Verlag, 2005, p. 46-58.

Gras R. et Kuntz P. (2008), An overview of the Statistical Implicative, *Statistical Implicative Analysis, R.Gras, E. Suzuki, F.Guillet and F.Spagnolo, Eds, Springer-Verlag, Berlin-Heidelberg,* p. 11-40.

Gras R. et Régnier J.C. (2009), Origine et développement de l'Analyse Statistique Implicative, *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* p. 6-16

Kodratoff Y. (2000), Extraction de connaissances à partir des données et des textes, *Actes du colloque ASI 2, IUFM de Caen.*

Guillaume S. (2002), Découverte de règles d'association ordinales, *EGC 2002, Volume 1,* Hermès, p. 29-40

Lagrange J.B. (1998), Analyse implicative d'un ensemble de variables numériques ; application au traitement d'un questionnaire aux réponses modales ordonnées, *Revue de Statistiques Appliquées, XLVI (1),* p. 71-93.

Lahanier-Reuter D. (1999), *Conceptions du hasard et enseignement des probabilités et statistiques*, Éducation et Formation, PUF.

Lallich S., Lenca p. et Vaillant B. (2004), Variation autour de l'intensité d'implication, *Actes de la Troisième Rencontre Internationale A.S.I., Università degli Studi di Palermo,* p. 237-246

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

Leblond J.-M., (1981), *L'esprit de sel*, Fayard

Lenca P., Meyer P., Vaillant B., Picouet P. et Lallich S. (2004), Évaluation et analyse multicritère des mesures de qualité des règles d'association, *Mesures de qualité pour la fouille de données, RNTI-E-1, Cépaduès –Éditions,* p 219-246

Lenca P., Vaillant B., Meyer P., et Lallich S. (2007), Association Rule Interestingness Measures : Experimental and Theorical Studies, *F.Guillet and H. J.Hamilton eds, Studies in Computational Intelligence 43, Springer*, p. 51-76.

Lerman I.C., Gras R., Rotsam H. (1981a), Élaboration 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. (1981b), *Classification et analyse ordinale des données*, Paris, Dunod

Lerman I.C. et Azé J. (2004), Indice probabiliste discriminant (de vraisemblance du lien) d'une règle d'association en cas de « très grosses données », *Mesures de qualité pour la fouille de données, RNTI-E-1, Cépaduès –Editions*, p 69-94.

Loevinger J. (1947), A systematical approach to the construction and evaluation of tests of ability, *Psychological Monographs*, n° 61, 1947, p. 1-49.

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

Ratsimba-Rajohn H. (2009), Guide d'utilisation des principales fonctionnalités du logiciel CHIC (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* p.295-315.

Ritschard G., Zighed D. et Marcellin S., (2007), Données déséquilibrées, entropie décentrée et indice d'implication, *Actes de la Quatrième Rencontre Internationale A.S.I, Universitat Jaume I, Castellon*, p.315-328.

Ritschard G., Marcellin S., Zighed D.A. (2009), Arbre de décision pour données déséquilibrées : sur la complémentarité de l'intensité d'implication et de l'entropie décentrée, *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* p.207-219.

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

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

Terano T., Liu H. et Chen L.P., eds, *Association Rules, volume 1805 of Lectures Notes in Computer Science, Spinger.*

Thom R., (1980), *Paraboles et catastrophes*, Flammarion, Paris.

Xuan-Hiep Huynh, Guillet F., Blanchard J., Kuntz P., Briand H. et Gras R. (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

Ouvrages de référence :

L'implication statistique. Nouvelle méthode exploratoire de donnée, sous la direction de R.Gras, et la collaboration de S. Ag Almouloud, M. Bailleul, A. Larher, M. Polo, H. Ratsimba-Rajohn, A.Totohasina, La Pensée Sauvage, Grenoble.

Mesures de Qualité pour la Fouille de Données, H.Briand, M.Sebag, R.Gras er F.Guillet eds, RNTI-E-1, Cépaduès, 2004

Quality Measures in Data Mining , F.Guillet et H.Hamilton eds, Springer, 2007,

Statistical Implicative Analysis, Theory and Applications, R.Gras, E. Suzuki, F. Guillet, F. Spagnolo, eds, Springer, 2008.

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, 2009.

Teoria y Aplicaciones del Analisis Estadistico Implicativo, Eds : P.Orus, L.Zemora, P.Gregori, Universitat Jaume-1, Castellon (Espagne), ISBN : 978-84-692-3925-4.

Recebido em 4/6/2013

Aceito em 4/7/2013