



## Variations autour de l'intensité d'implication

Stéphane Lallich\*, Philippe Lenca\*\*, Benoît Vaillant\*\*

\*Laboratoire ERIC, Université Lyon 2,  
 5 avenue Pierre Mendès France, 69676 Bron Cedex France  
 stephane.lallich@univ-lyon2.fr  
<http://eric.univ-lyon2.fr/~lallich/>

\*\* CNRS UMR 2872 TAMCIC, GET - ENST Bretagne  
 Technopôle de Brest Iroise, CS 83818, 29238 BREST Cedex  
 philippe.lenca@enst-bretagne.fr  
<http://perso.enst-bretagne.fr/~lenca/>  
 benoit.vaillant@enst-bretagne.fr  
<http://perso.enst-bretagne.fr/~vaillant/>

**Résumé.** Ce papier porte sur les mesures statistiques et probabilistes de l'intérêt des règles d'association, plus particulièrement sur celles issues de la modélisation de la loi du nombre de contre-exemples sous une hypothèse de référence, indépendance le plus souvent, mais aussi indétermination. Nous offrons d'abord une présentation synthétique de ces différentes mesures qui met en évidence les options de modélisation sous-jacentes, ce qui nous permet de proposer quelques variantes utiles. Dans un second temps, nous entreprenons de paramétrer l'hypothèse de référence par rapport à un seuil  $\theta$  laissé au choix de l'utilisateur. En suivant la logique précédemment décrite, nous proposons une gamme de mesures paramétrées qui généralisent les principales mesures construites autour de la loi du nombre de contre-exemples, indice d'implication, intensité d'implication, intensité d'implication entropique, indice probabiliste discriminant ou indice d'écart à l'équilibre.

### 1 Motivations

Dans ce papier, nous nous intéressons à l'évaluation objective de l'intérêt des règles d'association  $A \rightarrow B$  telles qu'elles ont été popularisées par (Agrawal et al. 1993, Agrawal et Srikant 1994), au moyen de mesures statistiques ou probabilistes. Dans de telles règles,  $A$  et  $B$  sont des conjonctions d'attributs binaires ou plus généralement des conjonctions de tests sur les valeurs prises par des attributs et l'on s'intéresse plus particulièrement aux « coprésences » de  $A$  et  $B$ . Les règles sont apprises à partir d'une base de données qui comporte  $n$  cas décrits par des attributs. On note  $n_a$  (resp.  $n_b$ ,  $n_{ab}$ ,  $n_{\bar{a}\bar{b}}$ ) le nombre de cas qui réalisent  $A$  (resp.

$B$ ,  $A$  et  $B$ ,  $A$  et  $\bar{B}$ ). Les fréquences relatives correspondantes sont notées  $p_a$  (resp.  $p_b$ ,  $p_{ab}$ ,  $p_{\bar{a}\bar{b}}$ ).

Le support  $p_{ab}$  et la confiance  $p_{b/a}$  fondent la stratégie utilisée dans les principaux algorithmes d'extraction, à la suite d'Apriori (Agrawal and Srikant 1994). Ces algorithmes sélectionnent l'ensemble des règles dont le support et la confiance dépassent les seuils de support et de confiance préalablement fixés,  $\sigma_s$  et  $\sigma_c$ . Ce type d'algorithme, exhaustif et déterministe (Freitas, 2000), produit des règles trop nombreuses dont l'intérêt n'est pas toujours assuré. On attend d'une règle intéressante que sa confiance dépasse une valeur de référence, mais celle-ci ne coïncide pas forcément avec  $\sigma_c$ , le seuil utilisé lors de l'extraction. Il est donc nécessaire de disposer d'autres mesures objectives. Ces mesures prennent en compte l'écart de la confiance à la valeur de référence, tout en le pondérant de telle sorte que diverses propriétés soient satisfaites, propriétés détaillées notamment dans (Hilderman et Hamilton 1999, Gras et al. 2004, Lallich et Teytaud 2004, Lenca et al. 2004). Bien souvent, les cas décrits dans la base de données ne sont qu'un échantillon. Il est alors fondé de construire des mesures dans lesquelles intervient la taille de l'échantillon. On distingue ainsi les mesures statistiques et les mesures descriptives. Pour un même tableau de fréquences relatives, une mesure statistique



croît avec  $n$ , la taille de l'échantillon, alors qu'une mesure descriptive reste inchangée. A la suite des travaux de (Gras 1979, Lerman et al. 1981, Gras et al. 2001, Lallich et al. 2005) on peut retenir le procédé suivant pour construire efficacement une mesure statistique. Après avoir choisi une grandeur d'intérêt, par exemple le nombre de contre-exemples de la règle, un modèle aléatoire et une hypothèse nulle  $H_0$  qui spécifie la valeur de référence pour la confiance de la règle, on détermine la loi de la grandeur d'intérêt sous  $H_0$ . On construit alors une mesure statistique en centrant et réduisant la grandeur d'intérêt sous  $H_0$  (par exemple, l'indice d'implication). On passe à une mesure dite probabiliste en calculant le complément à 1 de la  $p$ -value unilatérale de la mesure statistique (par exemple, l'intensité d'implication).

Le plus souvent, la valeur de référence utilisée pour la confiance est  $p_b$  qui correspond à l'indépendance de  $A$  et  $B$ ; la mesure évalue alors la dépendance positive entre  $A$  et  $B$ . Cependant, ainsi que le soulignent (Blanchard et al. 2005), la valeur de référence est parfois 0.5. Cette valeur correspond à la situation où il y a autant d'exemples que de contre-exemples, situation que l'on peut qualifier d'indétermination (en référence au vocabulaire statistique usuel) ou d'équilibre (ainsi (Guillet 2004)). Cette situation prévaut (Blanchard et al. 2005) pour des mesures descriptives comme la mesure de Sebag, la mesure de Ganascia, la moindre contradiction ou le taux d'exemples et de contre-exemples, ainsi que pour IPEE, l'indice probabiliste d'écart à l'équilibre qu'ils proposent. Plus généralement (Lallich et al. 2005), l'utilisateur peut être intéressé par une valeur de référence  $\theta$ , telle que  $0 < \sigma_c \leq \theta \leq 1$ . Par exemple, dans le cas où  $p_a = 0.05$  et  $p_b = 0.1$  (exemple section 3.3), la valeur de  $p_{b/a}$  sous l'indépendance est 0.1. Un utilisateur pourra choisir un seuil de 0.2 qui indique que  $B$  a au moins 2 fois plus de chances de se réaliser, sachant que  $A$  s'est réalisé. Ce seuil  $\theta$  peut aussi être celui à partir duquel une règle s'écarte significativement de l'indépendance, voire même être le seuil  $\sigma_c$  utilisé par l'algorithme d'extraction, comme c'est le cas pour le gain défini par (Fukuda et al. 2005). Concrètement, en médecine, celui qui utilise la règle pour l'aider à prendre une décision dépendant de l'état du conséquent  $B$  utilisera un seuil de 0.5 (prédiction), alors que celui qui cherche des facteurs de prédisposition (ciblage) fixera son seuil  $\theta$  en fonction du seuil d'indépendance  $p_b$ , par exemple  $\theta = 2p_b$ .

Notre contribution porte sur les mesures objectives de type statistique ou probabilistes. Après une présentation synthétique de l'approche utilisée pour construire les mesures statistiques et probabilistes usuelles (section 2), nous reprenons cette approche en paramétrant la valeur de référence de la confiance (section 3) et nous proposons différentes variantes paramétrées des mesures statistiques et probabilistes usuelles, plus particulièrement celles construites autour de l'indice d'implication, l'intensité d'implication, l'intensité d'implication entropique et l'indice probabiliste d'écart à l'équilibre.

## 2 Construction des indices statistiques et probabilistes usuels

### 2.1 Différentes modélisations

L'approche classique raisonne en référence à l'hypothèse d'indépendance de  $A$  et  $B$ . La modélisation de cette hypothèse d'indépendance a été étudiée par (Lerman et al. 1981), qui ont distingué trois modélisations différentes, comportant respectivement 1, 2 et 3 niveaux d'alea. Les étapes de ces différentes modélisations sont rassemblées dans le tableau 1. A la modélisation 1, nous avons associé une variante 1' où c'est  $p_a$  qui est fixé et non pas  $n_a$  (schéma binomial et non plus hypergéométrique).

Dans la suite, on note  $N_{ab}$  la variable aléatoire qui génère  $n_{ab}$ ,  $N_{\bar{a}\bar{b}}$  celle qui génère  $n_{\bar{a}\bar{b}}$ . Les notations *Hyp* et *Bin* désignent respectivement la loi hypergéométrique et la loi binomiale, alors que la loi de Poisson est désignée par *Poi* et la loi normale centrée réduite par  $N(0, 1)$ . Le principe des différents indices statistiques ou probabilistes construits à partir de  $n_{ab}$  et  $n_{\bar{a}\bar{b}}$  est le suivant. On établit la loi de  $N_{ab}$  et  $N_{\bar{a}\bar{b}}$  sous l'hypothèse nulle correspondant à la modélisation choisie, notée  $H_0$ . On en déduit un premier type d'indice statistique correspondant au nombre de contre-exemples de la règle, centré et réduit sous  $H_0$ , noté  $N_{ab}^{CR}$ . La loi exacte de cet indice peut être approximée par une loi normale, dans les conditions usuelles. On construit



alors un indice probabiliste, défini comme le complément à 1 de la probabilité d'observer sous  $H_0$  une valeur de l'indice statistique allant autant dans le sens de l'intérêt de la règle. La modélisation choisie n'a pas d'effet sur l'espérance, mais elle modifie la variance. La modélisation 3 est privilégiée par (Lerman et al. 1981), car elle traite les associations positives et les associations négatives de la façon la plus dissymétrique, respectant en cela l'esprit des règles d'association. Par centrage et réduction de  $N_{ab}$ , on obtient ainsi INDIMP, l'indice d'implication (Lerman et al. 1981), à partir duquel l'indice probabiliste formé est INTIMP, l'intensité d'implication (Gras 1979, Gras et al. 1996), que l'on peut considérer comme le complément à 1 de la  $p$ -value unilatérale du nombre de contre-exemples ou de l'indice d'implication. Cette mesure répond de façon satisfaisante à la plupart des critères de qualité usuels (Gras et al. 2004, Lenca et al. 2004).

## 2.2 Conservation du pouvoir discriminant

A côté de ses bonnes qualités, un inconvénient majeur de l'intensité d'implication, commun aux différentes mesures statistiques et probabilistes est sa perte de pouvoir discriminant. Par construction, elle évalue entre 0.95 et 1 les règles qui s'écartent significativement de l'indépendance ! Lorsque  $n$  est grand, en *data mining* tout particulièrement, le moindre écart à l'indépendance devient très significatif et la valeur des indices probabilistes reste collée à 1.

Pour corriger cette perte de pouvoir discriminant, (Lerman et Azé 2003) proposent une solution simple reposant sur une approche contextuelle où INDIMP est préalablement centré et réduit sur une base d'exemples  $\mathcal{B}$  (notation  $^{CR/B}$ ), définissant l'indice probabiliste discriminant IPD défini par :

$$IPD = P(N(0, 1) > INDIMP^{CR/B}).$$

	Modélisations 1, 1'	Modélisation 2	Modélisation 3
Principe	1.1 $n_a$ fixé $N_{ab}$ aléatoire pur  ou  1.1' $p_a$ fixé $N_{ab}$ aléatoire pur	2.1 $N_a \equiv Bin(n, p_a)$   2.2 $N_a = n_a$ $N_{ab} \equiv Bin(n_a, p_b)$	3.1 $N \equiv Poi(n)$ 3.2 $N = n$ , $N_a \equiv Bin(n, p_a)$ 3.3 $N = n$ , $N_a = n_a$ , $N_{ab} \equiv Bin(n_a, p_b)$
Loi $N_{ab}$ sous $H_0$	1.1 $Hyp(n, n_a, p_b)$ 1.1' $Bin(n_a, p_b)$	$Bin(n, p_a p_b)$	$Poi(np_a p_b)$
Loi $N_{a\bar{b}}$ sous $H_0$	1.1 $Hyp(n, n_a, p_{\bar{b}})$ 1.1' $Bin(n_a, p_{\bar{b}})$	$Bin(n, p_a p_{\bar{b}})$	$Poi(np_a p_{\bar{b}})$
Indice statistique $N_{ab}^{CR}$	1.1 $\frac{N_{ab} - np_a p_{\bar{b}}}{\sqrt{np_a p_{\bar{b}} p_b p_{\bar{b}}}}$ $= -r\sqrt{n}$ 1.1' $\frac{N_{ab} - np_a p_{\bar{b}}}{\sqrt{np_a p_b p_{\bar{b}}}}$	$\frac{N_{ab} - np_a p_{\bar{b}}}{\sqrt{np_a p_{\bar{b}} (1 - p_a p_{\bar{b}})}}$	INDIMP = $\frac{N_{ab} - np_a p_{\bar{b}}}{\sqrt{np_a p_{\bar{b}}}}$
Indice probabiliste $P(N(0,1) > N_{ab}^{cr})$	1.1 $P(N(0,1) < r)$		INTIMP = $P(N(0,1) > INDIMP)$

TAB 1 – Indices statistiques et probabilistes suivant différentes modélisations.



Une autre solution, faisant appel à une pondération, est proposée par (Gras et al. 2001) qui pondèrent INTIMP par la prise en compte d'un indice d'inclusion, fonction de l'entropie des expériences  $B/A$  et  $\bar{A}/\bar{B}$ . On note  $H(X) = p_x \log_2 p_x + p_x^- \log_2 p_x^-$  l'entropie de l'événement  $X$ . Suivant (Blanchard et al. 2004), la définition la plus générale de l'indice d'inclusion est :

$$I(A \subset B) = \left[ \left( 1 - H^*(B/A) \right) \times \left( 1 - H^*(\bar{A}/\bar{B}) \right) \right]^{1/2\alpha}$$

où  $H^*(X) = H(X)$ , si  $p_x > 0.5$  et  $H^*(X) = 1$ , sinon.

Le paramètre  $\alpha$  est choisi par l'utilisateur, la valeur  $\alpha = 2$  étant conseillée si l'on veut que l'indice soit tolérant aux premiers contre-exemples, ce que l'on fera dans la suite de ce texte. Par cette pondération, on obtient l'intensité d'implication entropique IIE définie par :

$$IIE = [\text{INTIMP} \times I(A \subset B)]^{1/2}.$$

La modulation de  $H(X)$  en  $H^*(X)$  a pour but d'éliminer des situations jugées sans intérêt, que  $p_{b/a} < 0.5$  (confiance de  $A \rightarrow B$  inférieure à 0.5) ou que  $p_{a/b}^- < 0.5$  (confiance de la contraposée  $\bar{B} \rightarrow \bar{A}$  inférieure à 0.5) ce qui correspond à une logique de prédiction. Il faut noter que dans une logique de ciblage, souvent utile en marketing ou en médecine, il serait logique de comparer  $p_{b/a}$  à  $p_b$  et  $p_{a/b}^-$  à  $p_a^-$  (Lallich et al. 2005).

Cette pondération de l'intensité d'implication par l'indice d'inclusion, pour efficace qu'elle soit, pose différents problèmes. L'indice d'inclusion est une mesure d'écart à l'indétermination fondée sur l'entropie qui s'annule naturellement lorsque  $p_{b/a} = 0.5$ , annulant la valeur de IIE. En revanche l'indépendance donne  $\text{INTIMP} = 0.5$  et n'annule pas IIE. C'est ainsi que sous l'indépendance, IIE n'est pas nulle et a une valeur qui dépend des

marges, à savoir  $\sqrt{\frac{(1-H(A))^2 \times (1-H(B))^2}{16}}$ , si à la fois  $p_a < 0.5$  et  $p_b > 0.5$ , et 0 sinon. Plus généralement,

en raisonnant sur le nombre de contre-exemples, lorsque  $p_a < 0.5$  et  $p_b > 0.5$ , les seuils d'indétermination de la règle  $A \rightarrow B$  (à savoir  $n_a/2$ ) et de sa contraposée ( $n_b^-/2$ ) sont au-dessus du seuil d'indépendance ( $n_a n_b^- / n$ ). Dans cette situation, IIE garde une valeur strictement positive, alors même que le nombre de contre-exemples est plus grand que celui attendu a priori, tant que  $n_a n_b^- / n \leq n_{ab}^- < \text{Min}\{n_a/2; n_b^-/2\}$ .

### 2.3 Intensité d'implication révisée

Pour supprimer les inconvénients signalés, nous proposons deux solutions qui autorisent chacune une nouvelle version de IIE, à savoir IIER (intensité d'implication entropique révisée) et IET (intensité d'implication entropique tronquée). La première solution consiste à remplacer INTIMP par  $\text{INTIMPR} = \text{Max}\{2\text{INTIMP}-1; 0\}$  dans la formule de IIE, obtenant  $\text{IIER} = [\text{INTIMPR} \times I(A \subset B)]^{1/2}$ . On supprime ainsi les inconvénients signalés, mais toutes les valeurs de IIE sont en théorie modifiées. En fait, cette modification est le plus souvent infime, puisque l'intensité d'implication est généralement très proche de 1.

La seconde solution se borne à annuler les valeurs de IIE pour les situations où le nombre de contre-exemples dépasse le seuil d'indépendance tout en étant inférieur aux seuils d'indétermination de la règle et de sa contraposée, soit  $n_a n_b^- / n \leq n_{ab}^- \leq \text{Min}\{n_a/2; n_b^-/2\}$ , sans changer les autres valeurs. Pour ce faire, on



modifie la modulation de  $H(X)$  comme suit. Une règle n'aura une évaluation non nulle au sens de l'indice d'inclusion que si tout à la fois :

- $p_{b/a} > 0.5$  (prédiction) et  $p_{b/a} > p_b$  (ciblage), soit  $p_{b/a} > \text{Max}\{0.5, p_b\}$
- $p_{a/b}^- > 0.5$  (prédiction) et  $p_{a/b}^- > p_a^-$  (ciblage), soit  $p_{a/b}^- > \text{Max}\{0.5, p_a^-\}$

On remplace donc la modulation  $H^*(X)$  qui précède (2.2) par  $H_t^*(X)$  défini comme suit :

- $H_t^*(B/A) = H(B/A)$ , si  $p_{b/a} > \text{Max}\{0.5, p_b\}$ ;  $H_t^*(B/A) = 1$ , sinon.
- $H_t^*(A/B) = H(A/B)$ , si  $p_{a/b}^- > \text{Max}\{0.5, p_a^-\}$ ;  $H_t^*(A/B) = 1$ , sinon

Après avoir choisi  $\alpha = 2$ , on en déduit l'indice d'inclusion tronqué et l'intensité d'implication entropique tronquée :

- $I_t(A \subset B) = \left[ \left( 1 - H_t^*(B/A) \right) \times \left( 1 - H_t^*(A/B) \right) \right]^{1/2\alpha}$
- $I_{IET} = [INTIMP \times I_t(A \subset B)]^{1/2}$

Ces deux variantes des indices usuels,  $I_t$  et  $I_{IET}$ , ont la propriété d'être nulles dès que le nombre de contre-exemples dépasse  $\text{Min}\{n_a n_b / n ; n_a / 2 ; n_b / 2\}$ , sans affecter les autres valeurs de ces indices.

#### 2.4 Mesure d'écart à l'indétermination

Après avoir remarqué que différentes mesures admettent 0.5 pour valeur de référence, (Blanchard et al. 2005) ont proposé IPEE, indice probabiliste d'écart à l'équilibre. Ces auteurs se placent implicitement dans le cadre de la modélisation 1' (tableau 1) puisqu'ils considèrent que l'on a  $N_{ab} \equiv \text{Bin}(n_a, 0.5)$  sous l'hypothèse

d'indétermination. Ils forment alors  $N_{ab}^{CR} = (N_{ab} - 0.5n_a) / 0.5\sqrt{n_a}$  pour proposer :

$$IPEE = P(\text{Bin}(n_a, 0.5) > n_{ab}) \approx P\left(N(0, 1) > \frac{n_{ab} - 0.5n_a}{0.5\sqrt{n_a}}\right)$$

Avec l'approximation normale, valide dès que  $n_a \geq 10$ , IPEE vaut 0.5 en cas d'indétermination. Cette mesure est clairement l'indice probabiliste associé à la modélisation 1' (tableau 1), où l'on a remplacé  $p_b$  par 0.5. En ce sens, elle hérite du caractère faiblement discriminant de ce type de mesure, sauf si 0.5 est nettement plus grand que  $p_b$ , ce qui amène ses promoteurs à suggérer de la pondérer par l'indice d'inclusion, ce qui est tout à fait cohérent, l'indice d'implication étant dans la même logique de référence à l'indétermination.

### 3 Paramétrisation des indices statistiques et probabilistes usuels

#### 3.1 Construction des indices statistiques et probabilistes généralisés

Comme nous l'avons fait pour les mesures descriptives, nous pouvons généraliser les mesures statistiques au cas où l'intérêt de la règle est évalué en comparant la confiance à un seuil  $\theta$ . Il suffit de reprendre notre tableau 1 et de considérer que pour chacune des modélisations proposées par (Lerman et al. 1981), la probabilité sous  $H_\theta$  d'un exemple de la règle, conditionnellement à  $n_a$ , est  $\theta$ , soit  $N_{ab} \equiv \text{Bin}(n_a, \theta)$ .

Le résultat de la modélisation 1 est immédiat, celui des modélisations 2 et 3 s'obtient facilement par les fonctions génératrices. On sait que si  $X \equiv \text{Bin}(m, p)$ , sa fonction génératrice s'écrit  $G(s) = E(s^X) = (1-p+ps)^m$  et que si  $X \equiv \text{Poi}(\lambda)$ , alors  $G(s) = E(s^X) = e^{-\lambda(1-s)}$ .

- modélisation 2 :  $n$  fixé,  $N_a \equiv \text{Bin}(n, p_a)$  et  $N_{ab}/N_a = n_a \equiv \text{Bin}(n_a, \theta)$   
alors  $G_{N_{ab}}(s) = E(s^{N_{ab}}) = E(E(s^{N_{ab}}/N_a)) = E((1-\theta + \theta s)^{N_a}) = (1-\theta p_a + \theta p_a s)^n$



Il vient :  $N_{ab} \equiv \text{Bin}(n, \theta p_a)$  et  $N_{a\bar{b}} \equiv \text{Bin}(n, (1-\theta)p_a)$

• modélisation 3 :  $N \equiv \text{Poi}(n)$ ,  $N_a/N=n \equiv \text{Bin}(n, p_a)$  et  $N_{ab}/(N=n \text{ et } N_a=n_a) \equiv \text{Bin}(n_a, \theta)$   
 alors  $G_{N_a}(s) = E(s^{N_a}) = E(E(s^{N_a}/N)) = E((1-p_a+p_a s)^N) = e^{-np_a(1-s)}$

Il vient  $N_a \equiv \text{Poi}(np_a)$

De même :  $G_{N_{ab}}(s) = E(s^{N_{ab}}) = E(E(s^{N_{ab}}/N_a)) = E((1-\theta+\theta s)^{N_a}) = e^{-n\theta p_a(1-s)}$

Soit :  $N_{ab} \equiv \text{Poi}(n\theta p_a)$  et  $N_{a\bar{b}} \equiv \text{Poi}(n(1-\theta)p_a)$

A partir de cette approche (tableau 2), nous proposons un cadre de modélisation et une gamme de mesures statistiques et probabilistes paramétrées suivant la valeur de référence à laquelle on compare la confiance, notamment l'indice d'implication généralisé  $\text{INDIMP}_{G_{|\theta}}$ , l'intensité d'implication généralisée  $\text{INDIMP}_{G_{|\theta}}$ , l'intensité d'implication entropique généralisée  $\text{IEG}_{|\theta}$ , qui correspondent à la modélisation 3 de (Lerman et al. 1981) et l'indice probabiliste d'écart généralisé  $\text{IPEG}_{|\theta}$  associé à la modélisation 1'.

	Modélisations 1, 1'	Modélisation 2	Modélisation 3
Principe	1.1 $n_a$ fixé $N_{ab}$ aléatoire pur  1.1' $p_a$ fixé $N_{ab}$ aléatoire pur	2.1 $N_a \equiv \text{Bin}(n, p_a)$  2.2 $N_a = n_a$ $N_{ab} \equiv \text{Bin}(n_a, \theta)$	3.1 $N \equiv \text{Poi}(n)$ 3.2 $N = n$ , $N_a \equiv \text{Bin}(n, p_a)$ 3.3 $N = n, N_a = n_a$ , $N_{ab} \equiv \text{Bin}(n_a, \theta)$
Loi $N_{ab}$ sous $H_0$	1.1 $\text{Hyp}(n, n_a, \theta)$ 1.1' $\text{Bin}(n_a, \theta)$	$\text{Bin}(n, p_a \theta)$	$\text{Poi}(np_a \theta)$
Loi $N_{a\bar{b}}$ sous $H_0$	1.1 $\text{Hyp}(n, n_a, 1-\theta)$ 1.1' $\text{Bin}(n_a, 1-\theta)$	$\text{Bin}(n, p_a(1-\theta))$	$\text{Poi}(np_a(1-\theta))$
Indice statistique $N_{ab}^{CR}$	1.1 $\frac{N_{a\bar{b}} - np_a(1-\theta)}{\sqrt{np_a p_a \theta(1-\theta)}}$  1.1' $\frac{N_{a\bar{b}} - np_a(1-\theta)}{\sqrt{np_a \theta(1-\theta)}}$	$\frac{N_{a\bar{b}} - np_a(1-\theta)}{\sqrt{np_a(1-\theta)(1-p_a(1-\theta))}}$	$\text{IndImp}_{G_{ \theta}} = \frac{N_{a\bar{b}} - np_a(1-\theta)}{\sqrt{np_a(1-\theta)}}$
Indice probabiliste $P(N(0,1) > N_{ab}^{cr})$	1.1 $\text{IPEG}_{ \theta}$		$\text{IntImp}_{G_{ \theta}} = P(N(0,1) > \text{IndImp}_{ \theta})$

TAB 2 – Construction des indices généralisés suivant différentes modélisations

### 3.2 Conservation du pouvoir discriminant des mesures généralisées

Les mesures généralisées statistiques ou probabilistes proposées ont un faible pouvoir discriminant. Pour améliorer celui-ci, nous suggérons deux solutions, l'une contextuelle à la manière de (Lerman et Azé 2003), l'autre par pondération à l'aide d'un indice d'inclusion suivant la voie tracée par (Gras et al. 2001).



Avec l'approche contextuelle, nous proposons de centrer et réduire INDIMP $G$  (ou son équivalent dans les modélisations 1 et 2) sur une base d'exemples  $\mathcal{B}$  et de former l'indice probabiliste discriminant généralisé défini par :

$$IPDG_{|\theta} = P(N(0, 1) > INDIMP_{G_{|\theta}}^{CR/\mathcal{B}}).$$

Dans le cadre de l'approche par pondération, on forme  $IIEG_{|\theta}$ , l'intensité d'implication entropique généralisée définie comme le produit de  $INTIMP_{G_{|\theta}}$  (ou son équivalent dans les modélisations 1 et 2) par un indice d'inclusion. Un souci de cohérence exige que cet indice d'inclusion ait  $\theta$  comme valeur de référence et non 0.5. A cet effet, Il faut transformer  $(1-H(B/A)^2)$  et  $(1-H(\bar{A}/\bar{B})^2)$  en fonctions de pénalisation qui s'annulent lorsque  $p_{b/a} = \theta$  et non pas lorsque  $p_{b/a} = 0.5$ .

En ce qui concerne  $H(B/A)$ , on lui substitue une fonction  $\tilde{H}_{|\theta}(B/A)$  obtenue en remplaçant  $p_{b/a}$  par :

$$\tilde{p}_{b/a} = \frac{p_{b/a}}{2\theta}, \text{ si } p_{b/a} \leq \theta; \tilde{p}_{b/a} = \frac{p_{b/a} + 1 - 2\theta}{2(1-\theta)}, \text{ si } p_{b/a} \geq \theta$$

S'agissant de  $H(\bar{A}/\bar{B})$ , on a deux possibilités :

- on choisit aussi  $\theta$  comme valeur de référence, option qui généralise l'indice d'inclusion proposé par (Gras et al. 2001) avec 0.5, ce qui amène à former  $\tilde{H}_{|\theta}(\bar{A}/\bar{B})$  obtenu en remplaçant dans  $H(\bar{A}/\bar{B})$  la quantité  $p_{a/\bar{b}}$  par :

$$\tilde{p}_{a/\bar{b}} = \frac{p_{a/\bar{b}}}{2\theta}, \text{ si } p_{a/\bar{b}} \leq \theta; \tilde{p}_{a/\bar{b}} = \frac{p_{a/\bar{b}} + 1 - 2\theta}{2(1-\theta)}, \text{ si } p_{a/\bar{b}} \geq \theta$$

- on choisit la valeur de référence naturellement associée à celle de  $\theta$ , soit  $1 - \frac{p_a}{p_b}(1-\theta)$ , sachant que

$$p_{a/\bar{b}} = 1 - \frac{p_a}{p_b}(1 - p_{b/a}); \text{ en cas de référence à l'indépendance, on a } \theta = p_b, \text{ ce qui donne}$$

logiquement  $p_a$  comme valeur de référence pour  $H_{|\theta}(\bar{A}/\bar{B})$ .

Il ne reste plus qu'à moduler ces fonctions en calculant  $\tilde{H}_{|\theta}^*(B/A)$  et  $\tilde{H}_{|\theta}^*(\bar{A}/\bar{B})$ , où :

$$\tilde{H}_{|\theta}^*(X) = \tilde{H}_{|\theta}(X), \text{ si } p_x > \theta; \tilde{H}_{|\theta}^*(X) = 1, \text{ sinon}$$

On en déduit un indice de pénalisation, notée  $I_{|\theta}$ , qui généralise l'indice d'inclusion usuel :

$$I_{|\theta} = \left[ \left( 1 - \tilde{H}_{|\theta}^*(B/A)^\alpha \right) \left( 1 - \tilde{H}_{|\theta}^*(\bar{A}/\bar{B})^\alpha \right) \right]^{1/2\alpha}$$

A partir de celui-ci, on forme l'intensité d'implication entropique généralisée, qui accroît le pouvoir discriminant de l'intensité d'implication généralisée  $INTIMP_{G_{|\theta}}$  :

$$IIEG_{|\theta} = [INTIMP_{G_{|\theta}} \times I_{|\theta}]^{1/2}$$



### 3.3 Illustration sur un exemple

Dans l'attente d'une expérimentation systématique de ces différentes variantes des indices statistiques et probabilistes usuels sur des jeux d'essai ou sur des données réelles, nous présentons ci-dessous quelques illustrations du comportement de certaines de ces variantes dans le cas où  $p_a = 0.05$ ,  $p_b = 0.10$ , en fonction de la valeur de la confiance  $p_{b/a}$ , situation dans laquelle on peut être aussi bien intéressé par une stratégie de ciblage que par une stratégie de prédiction. Nous avons retenu les seuils  $\theta = 0.10$  (ciblage : indépendance),  $\theta = 0.20$  (ciblage : B se produit 2 fois plus souvent lorsque A est réalisé) et  $\theta = 0.5$  (prédiction), pour lesquels, nous avons à chaque fois expérimenté deux variantes suivant la façon dont est fixé le seuil de  $H(\overline{A}/\overline{B})$  (notation  $\sim$  si le même  $\theta$  est utilisé,  $\sim\sim$  si on utilise le  $\theta$  naturellement associé). Ces graphiques sont cohérents avec la démarche proposée et montrent bien l'impact du paramètre retenu.

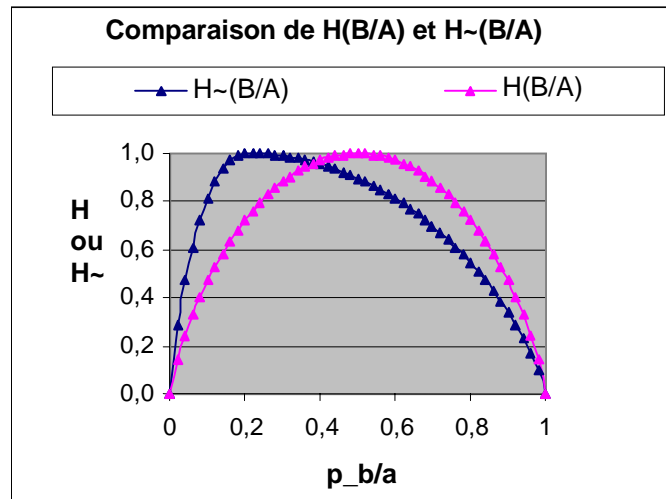


FIG 1 –  $H(B/A)$ , pour  $\theta = 0.50$  (prédiction, entropie usuelle) et  $\theta = 0.20$  (ciblage).

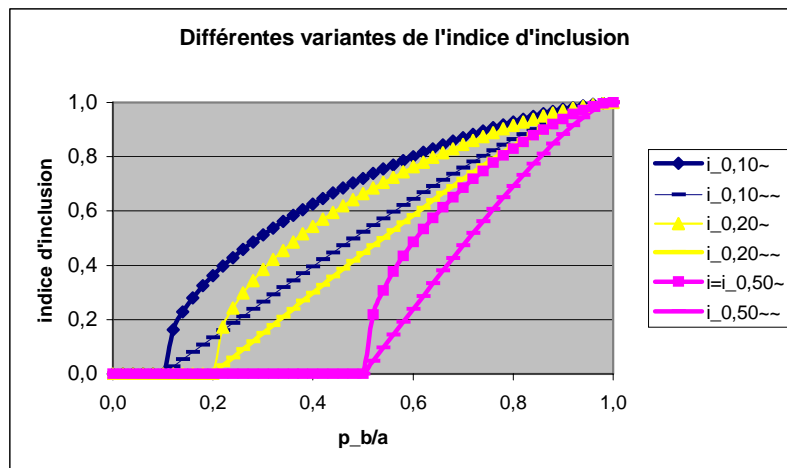




FIG 2 – Indice d'inclusion : indice usuel, variante  $\theta = 0.20$  (ciblage, notation  $\sim$  si le même  $\theta$  est utilisé pour la contraposée,  $\sim\sim$  si on utilise le  $\theta$  naturellement associé).

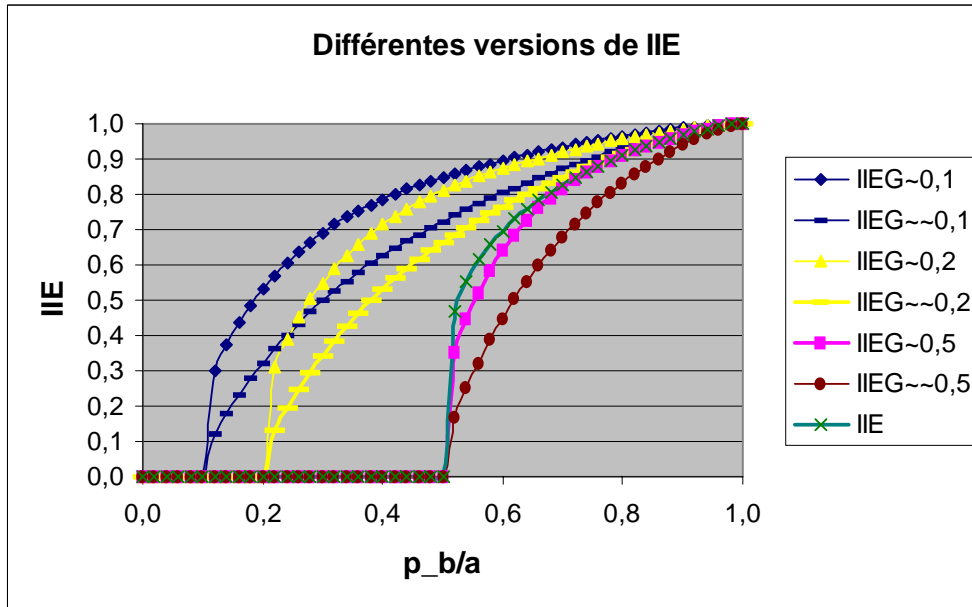


FIG 3 – Intensité d'implication entropique : IIE usuelle, 2 variantes  $\theta=0.50$  (prédiction), 2 variantes  $\theta=0.20$  (ciblage), 2 variantes  $\theta=0.10$  (indépendance)

#### 4 Conclusion

Guidés par un souci de cohérence et de modélisation, nous avons proposé dans cet article un cadre qui permet une présentation unifiée d'un grand nombre de mesures statistiques et probabilistes issues de la modélisation du nombre de contre-exemples des règles. Ce cadre aide à comprendre la logique de chacune de ces mesures et les liens qui existent entre elles. En outre, une telle présentation fonde de nouvelles mesures, indice d'implication généralisé, intensité d'implication généralisée, intensité d'implication entropique généralisée, indice probabiliste discriminant généralisé et indice probabiliste d'écart généralisé, qui font intervenir la valeur de référence de la confiance comme paramètre à la disposition de l'utilisateur.

#### Références

- Agrawal R. et Srikant R. (1994), Fast algorithms for mining association rules, J. B. Bocca, M. Jarke et C. Zaniolo, eds, Proceedings of the 20th Very Large Data Bases Conference, pp. 487-499, Morgan Kaufmann, 1994.
- Agrawal R., Imielinski T. et Swami A. (1993), Mining associations between sets of items in large databases, Proc. Of the ACM SIGMOD Conf., Washington DC, USA, 1993.
- Blanchard J., Guillet F., Briand H. et Gras R. (2005) IPEE : Indice probabiliste d'écart à l'équilibre pour l'évaluation de la qualité des règles, actes Atelier Qualité des Données et des Connaissances, pp. 26-34, 2005.
- Blanchard J., Kuntz P., Guillet F. et Gras R. (2004), Mesure de la qualité des règles d'association par l'intensité d'implication entropique. Revue des Nouvelles Technologies de l'Information (RNTI-E), 1, pp. 33-43, 2004.
- Freitas A. (1999), On rule interestingness measures, Knowledge-Based Systems Journal, pp. 309-315, 1999.



- Freitas A. (2000), Understanding the crucial differences between classification and discovery of association rules - a position paper, ACM SIGKDD Explorations, vol. 2, pp. 65-69, 2000.
- Fukuda T., Morimoto Y., Morishita S. et Tokuyama T. (1996), Data mining using two-dimensional optimized association rules : scheme, algorithms, and visualization, ACM SIGMOD International Conference on Management of Data, pp. 13-23, 1996.
- 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 Rennes 1, 1979.
- Gras R. et al. (1996), L'implication statistique, nouvelle méthode exploratoire des données, La Pensée Sauvage, Grenoble, 1996.
- Gras R., Couturier R., Bernadet M., Blanchard J., Briand H., Guillet F., Kuntz P., Lehn R. et Peter P. (2004), Quelques critères pour une mesure de qualité de règles d'association - un exemple : l'intensité d'implication. Revue des Nouvelles Technologies de l'Information (RNTI-E), 1, 2004.
- Gras R., Kuntz P., Couturier R. et Guillet F. (2001), Une version entropique de l'intensité d'implication pour les corpus volumineux, Actes 1e Conférence Extraction et Gestion des Connaissances, EGC 2001, Revue Extraction des connaissances et apprentissage, 1(1-2), pp. 69-80, 2001.
- Guillet F. (2004), Mesure de la qualité des connaissances en ECD, Tutoriel de la 4e Conf. Extraction et Gestion des Connaissances, EGC 2004, 60 p.
- Hilderman R. J. et Hamilton H. J. (1999), Knowledge discovery and interestingness measures : A survey, Technical Report 99-4, Department of Computer Science, University of Regina, oct. 1999.
- Lallich S. (2002), Mesure et validation en extraction des connaissances à partir des données. Habilitation à Diriger des Recherches, Université Lyon 2, 2002.
- Lallich S. et Teytaud O. (2004), Évaluation et validation de l'intérêt des règles d'association. Revue des Nouvelles Technologies de l'Information (RNTI-E), 1, pp. 193-217, 2004.
- Lallich S., Prudhomme E. et Teytaud O. (2004), Contrôle du risque multiple en sélection de règles d'association significatives, actes 4e Conférence Extraction et Gestion des Connaissances, EGC 2004, Revue des Nouvelles Technologies de l'information (RNTI-E-2), vol. 2, pp. 305-316.
- Lallich S., Vaillant B. et Lenca P. (2005), Parametrised measures for the evaluation of association rule interestingness, actes Conference "International Symposium on Applied Stochastic Models and Data Analysis", ASMDA 2005.
- Lenca P., Meyer P., Vaillant B., Picouet P. et Lallich S. (2004), Evaluation et analyse multi-critères des mesures de qualité des règles d'association, Revue des Nouvelles Technologies de l'Information (RNTI-E), 1, pp. 219-246, 2004.
- Lerman I.C. et Azé J. (2003), Une mesure probabiliste contextuelle discriminante de qualité des règles d'association, actes Conférence Extraction et Gestion des Connaissances, EGC 2003, RSTI-RIA, 1(17), pp. 247-262, 2003.
- Lerman I.C., Gras R. et Rostam H. (1981), Elaboration d'un indice d'implication pour les données binaires, i et ii, Revue Mathématiques et Sciences Humaines, n° 74, pp. 5-35, n° 75, pp. 5-47, 1981.
- Piatetsky-Shapiro G. (1991), Discovery, analysis and presentation of strong rules. In G. Piatetsky-Shapiro et W.J. Frawley, editors, Knowledge Discovery in Databases, pp. 229-248. AAAI/MIT Press, 1991.

## Summary

This paper deals with statistical and probabilistic association rule interestingness measures, particularly when considering the number of counter-example under a reference hypothesis, and its modelling. Our main concern in this paper is to propose an extension of the two classical references of independence and indetermination. First, we propose a synthetic overview of such classical measures, highlighting the underlying modelling possibilities. Through this synthesis, we were able to propose useful alternatives. Second, we parameterise the reference hypothesis using a threshold  $\theta$ , to be fixed by the user. With regards to this new formulation, we could generalise the considered set of interestingness measures, namely the implication index, the intensity of implication and its entropic extension, the probabilistic discriminant index and the probabilistic index of deviation from equilibrium.