Indice Probabiliste Discriminant de Vraisemblance du Lien pour des Données Volumineuses

Israël-César Lerman*, Jérôme Azé**

*Irisa-Université de Rennes 1
Campus de Beaulieu
35042 Rennes Cédex
lerman@irisa.fr
**Laboratoire de Recherche en Informatique
Université Paris-Sud
91405 Orsay
aze@lri.fr

Résumé. On sait que l'indice probabiliste implicatif usuel de vraisemblance du lien évaluant de façon intrinsèque une règle d'association, n'est plus discriminant si le nombre d'observations augmente suffisamment. Le but de cet article est de montrer l'extension discriminante de cet indice probabiliste pour évaluer une règle d'association, mais, dans le contexte d'un ensemble de règles. Cette approche a été proposée de longue date et a été validée dans le cadre de la classification hiérarchique AVL (Analyse de la Vraisemblance des Liens) d'un ensemble d'attributs de types quelconques. Une analyse expérimentale qui consiste à faire croître la taille des données par l'adjonction de contre-exemples à tous les attributs, montre toute la pertinence de la démarche statistique. Cette dernière est, au préalable, justifiée conceptuellement.

Mots Clés : Règle d'association, indice probabiliste discriminant, validité.

1 Introduction

La donnée est un tableau d'incidence ou d'existence à double entrée de dimension $n * p$ croisant un ensemble $\mathcal{O} = \{o_i | 1 \leq i \leq n\}$ de n objets avec un ensemble $\mathcal{A} = \{a^j | 1 \leq j \leq p\}$ de p attributs booléens. On peut noter α_i^j la valeur « vrai » ou « faux » de l'attribut a^j sur l'objet o_i : $\alpha_i^j = a^j(o_i), 1 \leq i \leq n, 1 \leq j \leq p$. On code généralement 1 la valeur « vrai » et 0, la valeur « faux ». On peut supposer sans restreindre la généralité que la valeur « vrai » à un attribut est sémantiquement plus signifiante que la valeur « faux » à cet attribut. Cela, le plus souvent, se traduit statistiquement par le fait que le nombre d'objets où l'attribut est à « vrai » est inférieur au nombre d'objets où l'attribut est à « faux ». La représentation que nous adoptons d'un attribut booléen a est son extension $\mathcal{O}(a)$ qui est le sous-ensemble des objets où a est à « vrai ». Ainsi $\mathcal{A}$ se trouve représenté par un ensemble de parties de l'ensemble

$\mathcal{O}$ des objets.
L'ensemble $\mathcal{O}$ des objets se trouve en général déterminé par un ensemble d'apprentissage ou échantillon d'un univers $\mathcal{U}$ d'objets. Quant à l'ensemble $\mathcal{A}$ des attributs booléens, il peut lui-même résulter de conjonctions (d'attributs booléens) provenant d'un ensemble $\mathcal{B}$. Ces dernières appelées communément « itemsets », sont récoltés dans le treillis des parties de $\mathcal{O}$ représentant les conjonctions de $\mathcal{B}$. Le problème de la détermination d'itemsets « significatifs » est un problème fondamental et à part entière de la fouille de données (« Data Mining »), mais qui ne nous concernera pas ici (Agrawal et al. 1993, Brin et al. 1997).

Ici, relativement à l'ensemble $\mathcal{A} * \mathcal{A}$ des couples d'attributs, l'objectif est de mettre en évidence ceux des couples de la forme $(a^j, a^k), 1 \leq j < k \leq p$, pour lesquels une valeur à « vrai » de l'attribut a^j est susceptible d'entraîner une valeur à « vrai » de l'attribut a^k. Il s'agit surtout et plutôt d'évaluer de façon relative les degrés d'implication de la forme $a^j \Rightarrow a^k$, au moyen d'un indice probabiliste de vraisemblance du lien, implicatif, valide sur les plans formel et statistique et discriminant en cas de très grosses données.
En effet, lorsqu'un tel indice a pu être proposé (Gras 1979, Lerman et al. 1981, Gras et Larher 1992) et, dans la mesure où une structure de liaison existait, il ne pouvait être discriminant que pour un nombre n d'objets qui ne soit pas trop élevé, disons inférieur à 10^3. Alors qu'il y a lieu maintenant de gérer de très grosses bases de données (n de l'ordre de centaines de milliers).
Relativement à un attribut a de A on désignera par $\neg a$ le sous ensemble complémentaire de $\mathcal{O}(a)$ dans $\mathcal{O}$. Relativement à un couple (a, b) de $\mathcal{A} * \mathcal{A}$, on introduit naturellement les conjonctions $a \wedge b$, $a \wedge \neg b$, $\neg a \wedge b$ et $\neg a \wedge \neg b$ qui sont respectivement représentées par $\mathcal{O}(a) \cap \mathcal{O}(b)$, $\mathcal{O}(a) \cap \mathcal{O}(\neg b)$, $\mathcal{O}(\neg a) \cap \mathcal{O}(b)$ et $\mathcal{O}(\neg a) \cap \mathcal{O}(\neg b)$. Les cardinaux respectifs de ces ensembles sont notés $n(a \wedge b)$, $n(a \wedge \neg b)$, $n(\neg a \wedge b)$ et $n(\neg a \wedge \neg b)$. Alors qu'on note $n(a)$ et $n(b)$ les cardinaux de $\mathcal{O}(a)$ et de $\mathcal{O}(b)$.

Aux précédents cardinaux, on peut respectivement associer les proportions qui rapportent ces cardinaux au nombre total n d'observations ($n = card(\mathcal{O})$). On a ainsi les proportions $p(a \wedge b)$, $p(a \wedge \neg b)$, $p(\neg a \wedge b)$ et $p(\neg a \wedge \neg b)$.

Alors qu'il s'agit de comparer deux à deux un ensemble d'implications valuées, l'indice probabiliste implicatif classique de vraisemblance du lien, mentionné ci-dessus, ne fait intervenir, relativement à l'évaluation $a \Rightarrow b$, que les paramètres liés à (a, b); soit $(n, p(a \wedge b), p(a \wedge \neg b), p(\neg a, b), p(\neg a, \neg b))$. Il s'agit, comme d'ailleurs pour ainsi dire, tous les indices proposés dans la littérature, d'une comparaison « absolue » ou « hors contexte », le contexte étant défini par l'ensemble de toutes les comparaisons. Cet indice que nous appellons aussi « local » peut s'apparenter au complément à 1 de ce que l'on appelle la « P-value » (seuil critique). L'optique est essentiellement distincte de celle des tests d'indépendance statistique (Bernard et Charron 1996). Elle se réfère plutôt à la philosophie de la théorie de l'information où il s'agit de quantifier un évènement qui nous intéresse par un indice qui se réfère à une échelle de probabilité. La valeur de cet indice est d'autant plus grande que l'évènement est orienté dans le sens que nous souhaitons, relativement à toutes les situations qu'il aurait pu occuper.

Dans notre cas, l'évènement concerné correspond à l'observation du tableau de contingence $\{n(a \wedge b), n(a \wedge \neg b), n(\neg a \wedge b), n(\neg a \wedge \neg b)\}$. Précisons de plus que le but n'est pas seulement, pour chacune des assertions $a^j \Rightarrow a^k$, de retenir un seuil de l'indice tel qu'au delà duquel l'implication se trouve validée. Il s'agit plutôt d'organiser de façon relative toutes les implications potentielles selon leurs intensités respectives.

Comme nous venons de l'exprimer, la plupart, sinon tous, des indices d'implication ont un caractère absolu et peuvent se réduire, notamment à des fins de comparaison entre implications, à des fonctions des paramètres de la forme $(p(a \wedge b), p(a \wedge \neg b), p(\neg a \wedge b), p(\neg a \wedge \neg b))$, qui sont des proportions. On pourra commencer par remarquer que ces paramètres se réduisent à un système de trois paramètres indépendants $(p(a \wedge b), p(a), p(b))$. Ces indices sont donc, par nature, discriminants quelle que soit la valeur n du nombre d'observations. Par contre, les valeurs des implications sont invariantes quel que soit n, pourvu que les proportions concernant les couples d'attributs, soient préservées.

Il en est ainsi par exemple des indices : Confiance (Agrawal et al. 1993), Lœvinger (Loevinger 1947), Piatetsky-Shapiro (Piatetsky-Shapiro 1991), Brin et al. (Brin et al. 1997), Corrélation (Pearson 1900), contribution orientée au χ^2 de Lerman et al. (Lerman et al. 1981), J-mesure de Goodman et Smith (Rodney M. Goodman 1988), Indice d'inclusion de Gras et al. (Gras et al. 2001). Cependant, l'évaluation de l'implication ne peut pas être insensible à la valeur de n.

Certains des indices précédents ont un caractère symétrique; c'est-à-dire, que leurs expressions sont invariantes si on remplace le couple (a, b) d'attributs booléens par le couple (b, a). Ils évaluent directement une similarité d'équivalence $(a \Leftrightarrow b)$, qui inclut bien sûr la similarité d'implication $(a \Rightarrow b)$. C'est surtout l'étude de la similarité d'équivalence qui a dominé la littérature de l'analyse des données.

Dans ces conditions, nous présenterons d'abord l'approche probabiliste de construction de l'indice de vraisemblance du lien entre attributs booléens dans le cas symétrique. Au paragraphe 2, on indique cette construction dans le cas absolu ou « hors contexte » où l'observation sur $\mathcal{O}$ du seul couple (a, b) d'attributs booléens est prise en compte. C'est la conception locale de l'indice probabiliste qui rend ce dernier non finement discriminant pour $n = card(\mathcal{O})$ suffisamment grand. C'est ainsi qu'au paragraphe 3, nous considérerons la construction de l'indice probabiliste de vraisemblance du lien « dans le contexte », relativement aux comparaisons mutuelles entre attributs de l'ensemble $\mathcal{A}$ des attributs. Nous montrerons comment nos constructions, d'abord « en dehors du contexte » puis « dans le contexte », rencontrent de façon originale des indices pouvant être reliés à la théorie du χ^2, notamment l'indice déjà mentionné, dit de la contribution orientée au χ^2. Au paragraphe 3, au cours de la construction, c'est une méthode de normalisation globale qui permet de situer de façon discriminante, les uns par rapport aux autres, les différents indices, dans le cadre d'une échelle de probabilité (Lerman 1984, Lerman et al. 1981). Le paragraphe 4 est dévolu à l'étude de la similarité implicative. On présente au paragraphe 4.1 quelques indices classiques

ou moins classiques en les analysant, qui s'expriment en fonction des seuls paramètres $(p(a \wedge b), p(a), p(b))$. Au paragraphe 4.2, on considère l'indice implicatif local de la vraisemblance du lien et celui dit d' « intensité entropique » (Gras et al. 2001). Enfin et surtout, nous développerons au paragraphe 4.3 la similarité implicative de vraisemblance du lien « dans le contexte », c'est-à-dire, relativement à un ensemble filtré d'implications potentielles. Cet indice, essentiellement probabiliste, qui suppose une normalisation globale par rapport au contexte implicatif, est nécessairement finement discriminant, quelle que soit la taille n de l'ensemble des objets. C'est précisément au paragraphe 5 que nous rapportons les résultats expérimentaux montrant le comportement du nouvel indice dit Intensité d'Implication Normalisée (IIN) relativement à l'indice plus classique de la vraisemblance du lien (Intensité d'Implication (II)). Une conclusion au paragraphe 6 précisera l'apport et la perspective qu'offre ce travail.

2 Comparaison « hors contexte » entre deux attributs

2.1 Démarche adoptée par rapport à une hypothèse d'absence de liaison (h.a.l.)

Soit (a, b) un couple d'attributs booléens issu de $\mathcal{A} * \mathcal{A}$. Nous avons déjà introduit ci-dessus la représentation ensembliste de ce dernier, ainsi que les paramètres cardinaux $n(a \wedge b)$, $n(a \wedge \neg b)$, $n(\neg a \wedge b)$ et $n(\neg a \wedge \neg b)$. Nous supposons pour fixer les idées que $n(a) < n(b)$. Le problème de l'évaluation du degré d'équivalence entre a et b $(a \Leftrightarrow b)$ a précédé celui du degré d'implication $a \Rightarrow b$. Pour le premier problème et dans l'optique de la comparaison deux à deux de l'ensemble $\mathcal{A}$ des attributs, de nombreux coefficients ont été proposés et qui peuvent tous s'exprimer en fonction des paramètres $(n(a \wedge b), n(a), n(b), n)$ ou d'ailleurs aussi en fonction des proportions $(p(a \wedge b), p(a), p(b))$ qui rapportent $n(a \wedge b)$, $n(a)$ et $n(b)$ à n. Ainsi, $n(a \wedge b)$ - qui représente le nombre d'objets où $a \wedge b$ est à « vrai » - apparaît comme un paramètre fondamental. Il s'agit de l'indice « brut » d'association. D'autre part, explicitement ou implicitement, la plupart des coefficients proposent chacun une normalisation par rapport aux tailles $n(a)$ et $n(b)$. En effet, toutes choses étant égales par ailleurs, une forte (resp. faible) valeur de ces paramètres conduit à une forte (resp. faible) valeur de $n(a \wedge b)$. Notre approche dans ces conditions a consisté à proposer une normalisation par rapport à un modèle probabiliste d'absence de liaison de la forme

$$(\mathcal{O}(a), \mathcal{O}(b), \mathcal{O}) \rightarrow (\mathcal{X}, \mathcal{Y}, \Omega) \tag{1}$$

où Ω est sinon $\mathcal{O}$, un ensemble aléatoire associé; et, pour Ω fixé, $\mathcal{X}$ et $\mathcal{Y}$ sont deux parties aléatoires indépendantes de Ω respectivement associées à $\mathcal{O}(a)$ et $\mathcal{O}(b)$. Le modèle que nous notons $\mathcal{N}$, est construit de telle façon que Ω, $\mathcal{X}$ et $\mathcal{Y}$ respectent sinon exactement, du moins en espérance mathématique les cardinaux n, $n(a)$, et $n(b)$. La partie aléatoire $\mathcal{X}$ (resp. $\mathcal{Y}$) peut également être notée $\mathcal{O}(a^*)$ (resp. $\mathcal{O}(b^*)$) où a^* (resp. b^*) est l'attribut booléen aléatoire associé à a (resp. b). Dans ces conditions, en

reprenant nos notations habituelles, à $s = n(a \wedge b) = card[\mathcal{O}(a) \cap \mathcal{O}(b)]$, nous associons sous le modèle (1), une variable aléatoire

$$\mathcal{S} = n(a^* \wedge b^*) = card(\mathcal{X} \cap \mathcal{Y}) \tag{2}$$

où (a^*, b^*) est le couple d'attributs booléens aléatoires indépendants correspondant à (a, b). $\mathcal{S}$ est l'indice brut aléatoire.
La première forme de normalisation consiste à centrer et à réduire s au moyen de l'espérance mathématique et de l'écart type de $\mathcal{S}$. On obtient dans ces conditions le coefficient

$$q(a, b) = \frac{s - \mathcal{E}(\mathcal{S})}{\sqrt{var(\mathcal{S})}} \tag{3}$$

où $\mathcal{E}(\mathcal{S})$ et $var(\mathcal{S})$ désignent l'espérance mathématique et la variance de $\mathcal{S}$.
L'indice probabiliste « local » de la vraisemblance du lien définit la deuxième forme de normalisation. Il s'écrit

$$\mathcal{I}(a, b) = Pr\{\mathcal{S} \leq s | \mathcal{N}\} = Pr\{q(a^*, b^*) \leq q(a, b) | \mathcal{N}\} \tag{4}$$

Dans un tel indice, le degré d'association entre a et b est évalué à partir du degré d'invraisemblance de la grandeur de s, eu égard à l'hypothèse d'absence de liaison $\mathcal{N}$. Bien que l'indice (4) corresponde au complément à l'unité d'un seuil critique au sens des tests d'hypothèse d'indépendance, il ne s'agit nullement ici d'un test conditionnel (Bernard et Charron 1996) mais d'une évaluation probabiliste conditionnée par les tailles $n(a)$ et $n(b)$.
Nous avons mis en évidence trois formes fondamentales de l'hypothèse d'absence de liaison $\mathcal{N}$ que nous notons $\mathcal{N}_1, \mathcal{N}_2$ et $\mathcal{N}_3$. Elles se distinguent dans la manière d'associer à un sous ensemble $\mathcal{O}(c)$ de $\mathcal{O}$, une partie aléatoire $\mathcal{L}$ d'un ensemble Ω (Lerman et al. 1981). Désignons ici par $P(\Omega)$ l'ensemble des parties de Ω organisé en niveaux à partir de la relation d'inclusion.
Pour $\mathcal{N}_1$, $\Omega = \mathcal{O}$ et $\mathcal{L}$ est un élément aléatoire pris uniformément au hasard sur le niveau $n(c)$ de $P(\Omega)$.
Pour $\mathcal{N}_2$, $\Omega = \mathcal{O}$; mais le modèle du choix $\mathcal{L}$ est à deux pas. Le premier consiste en le choix aléatoire d'un niveau et le second consiste en le choix aléatoire et uniformément réparti d'un élément de ce niveau. Précisons que le choix du niveau $k, 0 \leq k \leq n$, s'effectue avec la probabilité binomiale $C_n^k p(c)^k p(\neg c)^{n-k}$ où, avec des notations que l'on comprend, $p(c) = \frac{n(c)}{n}$ et $p(\neg c) = \frac{n(\neg c)}{n}$.
$\mathcal{N}_3$ est un modèle aléatoire à trois pas. Le premier consiste à associer à $\mathcal{O}$ un ensemble aléatoire Ω d'objets - pour fixer les idées, de même nature que ceux observés - la seule caractéristique aléatoire de Ω est sa cardinalité $\mathcal{N}$ qui est supposée suivre une loi de Poisson de paramètre $n = card(\mathcal{O})$. Les deux autres pas sont analogues à ceux du modèle $\mathcal{N}_2$. Pour $\mathcal{N} = m$ fixé et Ω_0 un ensemble de taille m, $\mathcal{L}$ est une partie aléatoire de Ω_0. $\mathcal{L}$ n'est définie que pour $m \geq n(c)$ et dans ce cas on pose γ le rapport $\frac{n(c)}{m}$.

Dans ces conditions, le choix du niveau k de $\mathcal{P}(\Omega_0)$ se fait avec la probabilité binomiale $C_m^k \gamma^k (1-\gamma)^{n-k}$. Pour un niveau choisi, le choix de $\mathcal{L}$ se fait alors uniformément au hasard sur ce niveau.
On démontre (Lerman et al. 1981) que la distribution de l'indice brut aléatoire $\mathcal{S}$ [*cf.* (2)] est :

- hypergéométrique de paramètres $(n, n(a), n(b))$ sous l'hypothèse $\mathcal{N}_1$;
- binomiale de paramètres $(n, p(a) * p(b))$ sous l'hypothèse $\mathcal{N}_2$;
- de Poisson de paramètres $(n, n * p(a) * p(b))$ sous l'hypothèse $\mathcal{N}_3$.

2.2 Les différentes formes d'un indice statistiquement normalisé

On démontre que, pour la première forme de normalisation, on obtient les indices suivants :

$$q_1(a,b) = \sqrt{n} * \frac{p(a \wedge b) - p(a) * p(b)}{\sqrt{p(a) * p(b) * p(\neg a) * p(\neg b)}} \tag{5}$$

$$q_2(a,b) = \sqrt{n} * \frac{p(a \wedge b) - p(a) * p(b)}{\sqrt{p(a) * p(b) * [1 - p(a) * p(b)]}} \tag{6}$$

et

$$q_3(a,b) = \sqrt{n} * \frac{p(a \wedge b) - p(a) * p(b)}{\sqrt{p(a) * p(b)}} \tag{7}$$

L'indice $q_1(a,b)$ est parfaitement symétrique dans ce sens où

$$q_1(a,b) = q_1(\neg a, \neg b) \tag{8}$$

On suppose que les attributs booléens (sous leur forme positive) ont été établis de telle façon que la proportion relative d'objets où un même attribut est à « vrai » est inférieure à 0.5 (voir l'introduction). Dans ces conditions, on peut démontrer les inégalités suivantes :

$$q_2(a,b) > q_2(\neg a, \neg b) \tag{9}$$

et

$$q_3(a,b) > q_3(\neg a, \neg b) \tag{10}$$

La dernière inégalité étant plus prononcée que celle qui précède, nous nous limiterons à considérer les deux indices les plus différenciés que sont $q_1(a,b)$ et $q_3(a,b)$. Une

autre raison de la rétention de ces deux indices est formelle et statistique. En désignant par $\chi^2\{(a,\neg a),(b,\neg b)\}$ la statistique du χ^2 associée au tableau de contingence $2*2$, croisant sur $\mathcal{O}$ $(a,\neg a)$ et $(b,\neg b)$, on a

$$\begin{aligned}\chi^2\{(a,\neg a),(b,\neg b)\} &= [q_1(a,b)]^2 \\ &= [q_3(a,b)]^2 + [q_3(a,\neg b)]^2 + \\ &\quad [q_3(\neg a,b)]^2 + [q_3(\neg a,\neg b)]^2 \end{aligned} \tag{11}$$

Ainsi, $q_3(a,b)$ est ce que nous appelons la contribution orientée de la case (a,b) à la statistique du χ^2.
En divisant par $\sqrt{n}$ les indices q_1 et q_3 on obtient les indices associés γ_1 et γ_3 qui ont le sens d'une corrélation, qui est comprise entre -1 et $+1$ pour γ_1 et dans un intervalle plus resserré pour γ_3 (dépendant de $p(a)$ et de $p(b)$) :

$$\gamma_1(a,b) = \frac{p(a\wedge b) - p(a)*p(b)}{\sqrt{p(a)*p(b)*p(\neg a)*p(\neg b)}} \tag{12}$$

et

$$\gamma_3(a,b) = \frac{p(a\wedge b) - p(a)*p(b)}{\sqrt{p(a)*p(b)}} \tag{13}$$

Désignons maintenant par d_{ab} la densité relative de la probabilité empirique jointe par rapport au produit des probabilités marginales. Nommément,

$$d_{ab} = \frac{p(a\wedge b)}{p(a)*p(b)} \tag{14}$$

Ce dernier indice est directement lié à celui t^{ab} introduit dans (Bernard et Charron 1996) au moyen de la relation

$$t^{ab} = d_{ab} - 1 \tag{15}$$

On peut maintenant exprimer $\gamma_1(a,b)$ et $\gamma_3(a,b)$ en fonction de d_{ab} et des proportions maginales. On a

$$\gamma_1(a,b) = \sqrt{\frac{p(a)*p(b)}{p(\neg a)*p(\neg b)}} * (d_{ab}-1) \tag{16}$$

$$\gamma_3(a,b) = \sqrt{p(a)*p(b)} * (d_{ab}-1) \tag{17}$$

2.3 Comportement de l'indice probabiliste local de la vraisemblance du lien

La forme générale de cet indice est donnée dans l'expression (4) où l'hypothèse d'absence de liaison $\mathcal{N}$ n'est pas spécifiée. En faisant respectivement $\mathcal{N} = \mathcal{N}_1$, $\mathcal{N} = \mathcal{N}_2$ ou $\mathcal{N} = \mathcal{N}_3$, on remplacera q par q_1, q_2 ou q_3. Grâce au calcul informatique, la probabilité $Pr(\mathcal{S} \leq s|\mathcal{N}_i)(i = 1, 2 \text{ ou } 3)$ peut être calculée de façon exacte pour n assez grand. On se réfèrera à la fonction de répartition de la loi hypergéométrique, binomiale ou de Poisson selon que $i = 1, 2$ ou 3. De toute façon, pour n assez grand (par exemple supérieur à 100) la loi de probabilité de $\mathcal{S}$ peut être approximée de façon précise par la loi normale et on a :

$$\mathcal{I}_i(a, b) = Pr\{\mathcal{S} \leq s|\mathcal{N}_i\} = \Phi[q_i(a, b)] = \Phi[\sqrt{n}\gamma_i(a, b)] \tag{18}$$

où Φ est la fonction de répartition de la loi normale centrée et réduite et où $i = 1, 2$ ou 3.
On se rend ainsi compte que si d_{ab} est sensiblement supérieur à 1, l'indice probabiliste local devient très voisin de l'unité pour n assez grand. D'autre part, si d_{ab} est sensiblement inférieur à 1, l'indice probabiliste devient très voisin de zéro pour n assez grand. Ainsi, si n est très grand, compte tenu de la précision calcul accessible, cet indice local ne permet - sur un ensemble $\mathcal{A}$ d'attributs booléens - que de distinguer : d'une part, les paires d'attributs liés positivement et d'autre part, les paires d'attributs liés négativement.
Imaginons maintenant, comme il est d'usage en inférence statistique, que l'ensemble $\mathcal{O}$ des objets est un ensemble d'apprentissage issu d'un univers $\mathcal{U}$ d'objets selon le mode du sondage aléatoire. Nous désignons respectivement par $\pi(a)$, $\pi(b)$ et $\pi(a \wedge b)$ les proportions d'objets où les attributs a, b et $a \wedge b$ sont à « vrai ». $\pi(a)$, $\pi(b)$ et $\pi(a \wedge b)$ s'interprètent également comme étant les probabilités que, respectivement a, b et $a \wedge b$ soient à « vrai » sur un objet pris uniformément au hasard dans $\mathcal{U}$. Désignons alors par $\rho_i(a, b)$ l'expression correspondante à $\gamma_i(a, b)$ au niveau de $\mathcal{U}$, $i = 1, 2$ ou 3. De la sorte, $\gamma_i(a, b)$ est une estimation de $\rho_i(a, b)$, d'autant plus précise que n est grand. Si on indique par $\delta(a, b)$ l'expression correspondant à (18) au niveau de $\mathcal{U}$, on comprend que, pour n suffisamment grand, si

- $\delta(a, b) < 1$, $\mathcal{I}_i(a, b)$ [*cf.* (18)] devient un nombre suffisamment voisin de 0 ;
- $\delta(a, b) > 1$, $\mathcal{I}_i(a, b)$ devient un nombre suffisamment voisin de 1 ;

alors que pour $\delta(a, b) = 1$, $\mathcal{I}_i(a, b)$ correspond à la réalisation d'une variable aléatoire uniformément répartie sur l'intervalle $[0, 1]$.
Toutefois, le contexte statistique dans lequel nous nous sommes situés est intrinsèque à l'ensemble $\mathcal{O}$ des objets et c'est dans ce contexte qu'il y a lieu d'aboutir à un indice probabiliste discriminant. Cependant, cet indice discriminant ne peut pas être obtenu si on n'avait à comparer et dans l'absolu, que deux attributs booléens a et b seulement. Et d'ailleurs, si notre univers se limitait à ces deux seuls attributs, l'indice $\mathcal{I}_i(a, b), \forall i = 1, 2$ ou 3, tel qu'il a été proposé [*cf.* (18)] est tout à fait satisfaisant. En fait, l'objectif consiste en la comparaison mutuelle de plusieurs paires d'attributs et généralement de l'ensemble $P_2(\mathcal{A})$ de toutes les paires d'attributs.

3 Comparaison « dans le contexte » entre deux attributs

Nous allons ici considérer l'évaluation relative d'une similarité entre attributs booléens dans le contexte d'un ensemble de paires d'attributs booléens. Il s'agira ici de l'ensemble des paires ou parties à deux éléments d'attributs d'un ensemble de p attributs booléens.

$$\mathcal{A} = \{a^j | 1 \leq j \leq p\} \tag{19}$$

Ainsi, à partir de $\mathcal{A} * \mathcal{A}$ on retient la structure cardinale suivante

$$\{(n(a^j \wedge a^k), n(a^j \wedge \neg a^k), n(\neg a^j \wedge a^k), n(\neg a^j \wedge \neg a^k)) | 1 \leq j < k \leq p\} \tag{20}$$

On notera également

$$\{n(a^j) | 1 \leq j \leq p\} \tag{21}$$

la suite des cardinaux des ensembles $\mathcal{O}(a^j), 1 \leq j \leq p$. En même temps que les précédents cardinaux, on peut introduire les proportions qui rapportent ces cardinaux au nombre total n d'observations. Aux tableaux (20) et (21) on associe les tableaux des proportions

$$\{(p(a^j \wedge a^k), p(a^j \wedge \neg a^k), p(\neg a^j \wedge a^k), p(\neg a^j \wedge \neg a^k)) | 1 \leq j < k \leq p\} \tag{22}$$

et

$$\{p(a^j) | 1 \leq j \leq p\} \tag{23}$$

Reprenons l'indice $q_3(a, b)$ [*cf.* (7)] défini localement par centrage et réduction relativement à l'hypothèse d'absence de liaison $\mathcal{N}_3$. Nous avons déjà fait remarquer que cet indice a une propriété intéressante de dissymétrie où la ressemblance entre attributs rares se trouve ponctuée [*cf.* (10)]. On suppose précisément que l'ensemble des p attributs booléens est établi de telle façon que, avec des notations que l'on comprend,

$$n(a^j) \leq n(\neg a^j), 1 \leq j \leq p \tag{24}$$

Nous avons déjà évoqué dans l'introduction, relativement aux indices invariants par rapport aux proportions qu'un même système de valeurs (22) ne peut pas être évalué de la même façon, relativement aux équivalences ou implications qu'il induit, quelle que soit la valeur de n, sur le plan statistique bien sûr, mais aussi sur le plan sémantique. Il s'agit de plus de situer de façon relative les différentes associations.

Nous avons également exprimé [*cf.* (11)] que $q_3(a,b)$ est la part orientée sur (a,b), relativement au χ^2 du tableau de contingence $2*2$, $\{a,\neg a\}*\{b,\neg b\}$.
Maintenant, pour comparer de façon mutuelle et relative l'ensemble des paires d'attributs, introduisons la variance empirique de l'indice q_3 sur l'ensemble $P_2(\mathcal{A})$ des parties à deux éléments de $\mathcal{A}$. Cette variance se met sous la forme:

$$var_e(q_3) = \frac{2}{p*(p-1)} \sum \{[q_3(a^j,a^k) - moy_e(q_3)]^2 | 1 \leq j < k \leq p\} \tag{25}$$

où

$$moy_e(q_3) = \frac{2}{p*(p-1)} \sum \{q_3(a^j,a^k) | 1 \leq j < k \leq p\} \tag{26}$$

représente la moyenne de l'indice q_3 sur $P_2(\mathcal{A})$.
Pour effectuer la comparaison relative entre deux attributs faisant partie de $\mathcal{A}$, a^1 et a^2 pour fixer les idées, on introduit l'indice $q_3^g(a^1,a^2)$, globalement normalisé. Il s'agit très précisément de la contribution relative et orientée de $q_3(a^1,a^2)$ à la variance $var_e(q_3)$. Ainsi, $q_3^g(a^1,a^2)$ a l'expression suivante:

$$q_3^g(a^1,a^2) = \frac{q_3(a^1,a^2) - moy_e(q_3)}{\sqrt{var_e(q_3)}} \tag{27}$$

L'indice probabiliste de la vraisemblance du lien se conçoit alors dans le cadre d'une hypothèse d'absence de liaison globale où à l'ensemble $\mathcal{A}$ des attributs [*cf.* (19)], on associe un ensemble

$$\mathcal{A}^* = \{a^{j*} | 1 \leq j \leq p\} \tag{28}$$

d'attributs aléatoires indépendants conformément à l'hypothèse d'absence de liaison $\mathcal{N}_3$. Cet indice s'écrit pour (a^1,a^2)

$$P_g(a^1,a^2) = Pr\{q_3^g(a^{1*},a^{2*}) \leq q_3^g(a^1,a^2) | \mathcal{N}_3\} \tag{29}$$

où

$$q_3^g(a^{1*},a^{2*}) = \frac{q_3(a^{1*},a^{2*}) - moy_e(q_3^*)}{\sqrt{var_e(q_3^*)}} \tag{30}$$

On démontre, et on s'en rend compte expérimentalement (Lerman 1984, Daudé 1992), que l'indice probabiliste (29) peut être calculé au moyen de

$$P_g(a^1,a^2) = \Phi[q_3^g(a^1,a^2)] \tag{31}$$

où Φ est la fonction de répartition de la loi normale centrée et réduite.
Signalons que la table des indices

$$\{P_g(a^j, a^k) | 1 \leq j < k \leq p\} \tag{32}$$

constitue l'argument de la méthode de classification ascendante hiérarchique AVL (Analyse de la Vraisemblance des Liens) (Lerman 1991). D'autre part, pour toute méthode d'analyse des données travaillant avec des dissimilarités, notre approche fournira la table des indices que nous appelons de « dissimilarité informationnelle » :

$$\{\mathcal{D}(j,k) = -Log_2(P_g(a^j, a^k)] | 1 \leq j < k \leq p\} \tag{33}$$

4 Indice de similarité implicative

4.1 Des indices indépendants de n et du contexte

Jusqu'à présent nous nous sommes intéressés à l'évaluation du degré d'équivalence ou d'implication mutuelle entre deux attributs booléens a et b, qu'on peut supposer issus de l'ensemble $\mathcal{A}$ des attributs [*cf.* (19)]. Nous nous sommes pour cela situés - via l'hypothèse d'absence de liaison - par rapport à l'hypothèse d'indépendance statistique. Les indices que nous considérerons ici se placent implicitement ou explicitement par rapport à cette hypothèse. Cependant, il s'agit ici d'évaluer une relation dissymétrique de la forme $a \Rightarrow b$. Une telle relation se trouve complètement vérifiée au niveau de l'ensemble $\mathcal{O}$ d'objets, si le sous-ensemble $\mathcal{O}(a)$ où a est à « vrai », se trouve inclus dans le sous ensemble $\mathcal{O}(b)$ où b est à « vrai ». En pratique, une telle circonstance est rare ; alors que la propension que b soit a « vrai » sachant que a est à « vrai », est relativement forte.

Cependant, la situation d'inclusion totale peut être considérée pour elle-même avec intérêt. On peut en effet, pour un indice donné, considérer la famille de cas paramétrée par $(n, p(a), p(b))$ où $p(a \wedge \neg b) = 0$. On peut aussi, à la manière de (Piatetsky-Shapiro 1991), s'intéresser à certains aspects de la loi de probabilité de la proportion conditionnelle $(p(b^*|a^*)$; mais sous un modèle aléatoire particulier où on suppose connaître la taille $N(a)$ du sous-ensemble de l'univers $\mathcal{U}$ où a est à « vrai », ainsi que la taille $n(a)$ des objets où a est à « vrai » dans l'échantillon $\mathcal{O}$. Les tailles de $\mathcal{O}$ et de $\mathcal{U}$ sont également supposées connues.

Maintenant, considérons le cas le plus réaliste et le plus fréquent où le nombre d'objets $n(a \wedge \neg b)$ est « petit » sans être nul.

Dans ces conditions, on se trouve conduit à s'intéresser à la petitesse relative du nombre d'objets $n(a \wedge \neg b)$ où a est à « vrai » sans que b le soit, c'est-à-dire, les objets qui contredisent la relation $a \Rightarrow b$. Pour évaluer de façon pertinente cette petitesse, la plupart des indices cherchent à neutraliser, chacun d'une façon, les effets des tailles $n(a)$ et $n(b)$. Cependant, pour $n(a)$ et $n(b)$ fixés, il y a une certaine concomittance des

phénomènes : $n(a \wedge \neg b)$ « petit », $n(a \wedge b)$ « grand », $n(\neg a \wedge b)$ « petit » et $n(\neg a, \neg b)$ « grand ».

De sorte que certains indices proposés ont un caractère parfaitement symétrique relativement au couple (a, b). Notre démarche s'inscrivant dans le contexte des comparaisons mutuelles, on peut neutraliser la valeur de n et s'exprimer au niveau déjà introduit des proportions $(p(a \wedge b), p(a \wedge \overline{b}), p(\overline{a} \wedge b), p(\overline{a} \wedge \overline{b}))$. La plupart, sinon tous, des indices peuvent se situer par rapport à l'hypothèse d'indépendance définie par $p(a \wedge b) = p(a) * p(b)$. Le plus simple des indices selon (Piatetsky-Shapiro 1991) est celui $(p(a \wedge b) - p(a) * p(b))$ qui correspond au numérateur commun de $q_1(a, b)$, $q_2(a, b)$ et $q_3(a, b)$ (cf. 5,6,7). Cet indice ainsi que celui $\gamma_1(a, b)$ sont présentés dans (Piatetsky-Shapiro 1991) au niveau de l'univers $\mathcal{U}$ des objets de taille N. Le dernier indice de corrélation γ_1 correspond à celui de K. Pearson (Pearson 1900). Dans (Brin et al. 1997), on présente également sous l'appellation « mesure d'intérêt » un indice symétrique, celui que nous avons appelé d_{ab} (cf. 14). Il est directement lié à la contribution au χ^2 de la cellule (a, b). Cette contribution peut s'exprimer par $[q_3(a, b)]^2$, alors, que ce que nous considérions dans (Lerman et al. 1981) c'est la contribution orientée $q_3(\alpha, \beta)$ pour chacune des cellules (α, β) de $\{a, \neg a\} \times \{b, \neg b\}$ (cf 11). Maintenant, l'objectif principal dans (Brin et al. 1997) est de pouvoir, grâce à une propriété de fermeture, procéder à l'élagage du treillis des itemsets, en utilisant l'indice du χ^2, plutôt que celui dit de confiance (cf. ci-dessous), comme il est usuel de le faire. Il en résulte un accent particulier et intéressant quant à la prise en compte des attributs sous la forme négative.

Néanmoins, la plupart des indices sont dissymétriques et ponctuent plus clairement le degré de petitesse de $n(a \wedge \neg b)$ comparativement à l'importance de $n(a \wedge b)$. Bien que cela n'intervienne pas directement, nous supposerons la condition de cohérence $n(a) \leq n(b)$ qui autorise l'inclusion totale de $\mathcal{O}(a)$ dans $\mathcal{O}(b)$.

L'indice le plus simple et le plus direct qu'il y a lieu d'évoquer est celui de la confiance (Agrawal et al. 1993). Il est défini par la proportion conditionnelle $p(b|a) = \frac{p(a \wedge b)}{p(a)}$. Il varie entre 0 et 1 ; 0 en cas de disjonction entre $\mathcal{O}(a)$ et $\mathcal{O}(b)$ et 1, en cas d'inclusion de $\mathcal{O}(a)$ dans $\mathcal{O}(b)$. Une analyse intéressante de cet indice, comparativement aux différents indices proposés dans la littérature est menée dans (Lallich et Teytaud 2003). Le deuxième indice très classique et qui fait référence qu'il y a lieu d'exprimer, est dû à J. Loevinger (Loevinger 1947). Il peut, par rapport à nos notations [*cf.* (14)] s'exprimer comme suit :

$$\mathcal{H}(a, b) = 1 - d(a, \neg b) \tag{34}$$

On suppose ici les deux inégalités « naturelles » $p(a) \leq p(b)$ et $p(a) \leq p(\neg b)$. Dans ces conditions, l'indice varie entre 1 en cas de l'inclusion totale $\mathcal{O}(a) \subset \mathcal{O}(b)$, passe par la valeur 0 au niveau de l'indépendance et atteint la valeur négative $-[\frac{p(b)}{p(\neg b)}]$ dans le cas où $\mathcal{O}(a) \subset \mathcal{O}(\neg b)$; c'est-à-dire, où c'est l'implication opposée à $a \Rightarrow \neg b$ qui se trouve vérifiée.

L'indice de Lœvinger peut aussi se mettre sous la forme:

$$\mathcal{H}(a,b) = \frac{p(a \wedge b) - p(a)p(b)}{p(a)p(\neg b)} \tag{35}$$

Il correspond donc à une réduction dissymétrique par rapport à (a,b) du premier indice proposé par G. Piatetsky-Shapiro (Piatetsky-Shapiro 1991).

On peut aussi clairement situer l'indice $\mathcal{H}(a,b)$ par rapport à celui $\gamma_3(a,\neg b)$. Alors que dans $\mathcal{H}(a,b)$, l'indice centré $[p(a \wedge \neg b) - p(a) * p(\neg b)]$ est réduit au moyen de $p(a) * p(\neg b)$, ce même indice centré est réduit au moyen de $\sqrt{p(a) * p(\neg b)}$ dans le coefficient $\gamma_3(a,\neg b)$. Ainsi, $-\gamma_3(a,\neg b)$, peut également jouer le rôle d'un indice d'implication, fonction seulement de $[p(a \wedge b), p(a), p(b)]$. Cet indice passe bien par la valeur 0 à l'indépendance, est égal à $\sqrt{p(a) * p(\neg b)}$ dans le cas où $a \Rightarrow b$ est complètement vérifiée et peut décroître jusqu'à $-p(b) * \sqrt{\frac{p(a)}{p(\neg b)}}$.

Dans ces conditions et clairement, l'indice discriminant que nous proposons, mais hors contexte, est défini par

$$-\gamma_3(a,\neg b) = \frac{-p(a \wedge \neg b) + p(a) * p(\neg b)}{\sqrt{p(a) * p(b)}} \tag{36}$$

il s'agit de l'opposé de la contribution orientée de la case $(a,\neg b)$ au coefficient χ^2/n. L'indice $-\gamma_3(a,\neg b)$ est cohérent avec la construction dans le contexte de l'indice probabiliste de vraisemblance du lien (voir paragraphe 4.3). On démontre que les deux bornes restent enserrées dans l'intervalle $[-1,+1]$; alors que dans le cas de $\mathcal{H}(a,b)$, la borne négative peut être, en valeur absolue, théoriquement aussi grande qu'on le veut. Plus précisément et relativement à l'indice $-\gamma_3(a,\neg b)$, on établit, sous les inégalités posées ($p(a) \leq p(b)$ et $p(a) \leq p(\neg b)$) que ses deux bornes sont comprises dans l'intervalle $[-p(b), p(\neg b)]$.

Si on considère la condition de cohérence $p(a) \leq p(b)$ permettant l'inclusion $\mathcal{O}(a) \subset \mathcal{O}(b)$ et si on suppose, comme c'est le cas général dans l'élaboration de l'ensemble des attributs booléens, que $p(a) \leq p(\neg a)$ et $p(b) \leq p(\neg b)$ (voir Introduction), alors on peut établir que les valeurs minimale et maximale de $\gamma_3(a,\neg b)$ sont, respectivement, -0.5 et +0.5. La borne supérieure ci-dessus $p(\neg b)$ est ainsi réduite. De sorte que, dans les conditions de cohérence mentionnées, si on désire un indice compris entre 0 et 1, on posera:

$$\eta_3(a,b) = 0.5 - \gamma_3(a,\neg b) \tag{37}$$

Signalons que dans les conditions de cohérence que nous venons de poser, la borne minimale pour l'indice $\mathcal{H}(a,b)$ est supérieure ou égale à -1. De sorte, qu'on peut alors définir un indice $\mathcal{K}(a,b)$ dérivé de $\mathcal{H}(a,b)$ et compris entre 0 et 1 sous la forme:

$$\mathcal{K}(a,b) = \frac{1}{2}(\mathcal{H}(a,b) + 1) \tag{38}$$

Les deux nouveaux indices deviennent égaux à $\frac{1}{2}$ à l'indépendance.
Nous avons vu que différents indices présentés ci-dessous pouvaient être situés par rapport au différents éléments rentrant dans la composition de la statistique du χ^2 associé au tableau de contingences 2×2, défini pour le croisement $\{a, \neg a\}\times\{b, \neg b\}$. Nous allons maintenant présenter deux indices qui utilisent les composants de la quantité d'information mutuelle associée à un tel tableau. Cette dernière, peut prendre l'une des trois formes suivantes :

$$\begin{aligned}
\mathcal{E} &= p(a\wedge b)\log_2(d(a,b)) + p(a\wedge\neg b)\log_2(d(a,\neg b)) \\
&+ p(\neg a\wedge b)\log_2(d(\neg a,b)) + p(\neg a\wedge\neg b)\log_2(d(\neg a,\neg b)) && (39)\\
&= E(a) - p(b)E(a|b) - p(\overline{b})E(a|\overline{b}) && (40)\\
&= E(b) - p(a)E(b|a) - p(\overline{a})E(b|\overline{a}) && (41)\\
& && (42)
\end{aligned}$$

où $E(x)$ représente l'entropie de la distribution $(p(x), p(\neg x))$ et où $E(x|y)$ représente celle de la distribution conditionnelle $(p(x|y), p(\neg x|y))$, x et y étant deux attributs booléens.
La J-mesure de Goodman & Smith (Rodney M. Goodman 1988) correspond précisément, dans la première expression (39), à la somme des deux premiers termes, où c'est l'attribut a qui est décliné. Alors que c'est l'attribut $\neg a$ qui est pris en considération dans la somme des deux derniers termes. Un deuxième indice qui a un caractère essentiellement entropique est dû à R. Gras (Gras et al. 2001) qui l'appelle indice d' « inclusion » Ce dernier prend la forme :

$$\tau(a,b) = \sqrt{G(b|a) * G(\neg a|\neg b)} \tag{43}$$

où $G(x|y)$ est la racine carrée positive de

$$G^2(x|y) = \begin{cases} 1 - E^2(x|y) \text{ si } p(\neg x\wedge y) \leq \frac{1}{2} * p(y) \\ 0 \text{ sinon} \end{cases} \tag{44}$$

Cet indice utilise les entropies conditionnelles $E(b|a)$ et $E(\neg a|\neg b)$ qui sont, respectivement, celles des distributions à deux valeurs $(p(b|a), p(\neg b|a))$ et $(p(\neg a|\neg b), p(a|\neg b))$. La seule première entropie peut être récoltée comme un composant constitutif de la dernière expression (41) de l'information mutuelle $\mathcal{E}$, alors que la deuxième entropie est un élément composant de la précédente expression (40) de $\mathcal{E}$.

On notera que l'importance de l'indice $1 - E^2(b|a)$ traduit tout autant l'inclusion $\mathcal{O}(a) \subset \mathcal{O}(b)$ que celle, logiquement contraire, $\mathcal{O}(a) \subset \mathcal{O}(\neg b)$ $(E(b|a) = E(\neg b|a))$. D'autre part, l'indice $1 - E^2(\neg a|\neg b)$, traduit tout autant l'inclusion $\mathcal{O}(\neg b) \subset \mathcal{O}(\neg a)$ que celle $\mathcal{O}(\neg b) \subset \mathcal{O}(a)$ $(E(\neg a|\neg b) = E(a|\neg b))$. Cependant, compte tenu de la condition apparaissant dans l'équation (44), une valeur positive et non nulle de $\tau(a,b)$ est conditionnée par $p(b|a) > p(\neg b|a)$. Ainsi, l'indice d'inclusion ne prend une valeur positive et non nulle que si chacun des deux supports $p(b|a)$ et $p(\neg a|\neg b)$ est supérieur à

0.5. Or, il y a des situations, peut-être non fréquentes, où $p(b|a)$ est suffisamment élevé (très supérieur à 0.5), alors que $p(\neg a|\neg b)$ est suffisamment bas (très inférieur à 0.5). Il est difficile dans ce cas de rejeter toute valeur à l'implication $a \Rightarrow b$. De sorte que l'indice d'inclusion possède la faiblesse de sa qualité, à savoir tenir compte de l'implication $a \Rightarrow b$ et de sa contraposée $\neg b \Rightarrow \neg a$.

4.2 Indice implicatif local de la vraisemblance du lien et intensité entropique

Maintenant, dans la conception de l'indice probabiliste local de la vraisemblance du lien pour mesurer la similarité entre deux attributs a et b, nous cherchons à évaluer le degré d'invraisemblance de la grandeur $n(a \wedge b)$. Relativement à l'évaluation de l'implication $a \Rightarrow b$, l'idée de R. Gras (Gras 1979) a consisté à transporter la démarche pour se poser la question du degré d'invraisemblance de la petitesse de $n(a \wedge \neg b)$, eu égard à l'hypothèse d'absence de liaison $\mathcal{N}$ qui neutralise d'une certaine façon les influences de $n(a)$ et de $n(b)$, l'indice prend la forme

$$\begin{aligned} \mathcal{I}(a,b) &= 1 - Pr\{n(a^* \wedge \neg b^*) < n(a \wedge \neg b)|\mathcal{N}\} \\ &= Pr\{n(a^* \wedge \neg b^*) \geq n(a \wedge \neg b)|\mathcal{N}\} \end{aligned} \quad (45)$$

où (a^*, b^*) est le couple d'attributs aléatoires indépendants associé à (a, b) dans l'hypothèse $\mathcal{N}$.
Désignons ici pas u l'indice $n(a \wedge \neg b)$ et par u^*, l'indice aléatoire $n(a^* \wedge \neg b^*)$ considéré dans l'hypothèse d'absence de liaison $\mathcal{N}$.
L'indice centré et réduit prend la forme

$$q(a, \neg b) = \frac{u - \mathcal{E}(u^*)}{\sqrt{var(u^*)}} \quad (46)$$

où $\mathcal{E}(u^*)$ et $var(u^*)$ sont respectivement l'espérance mathématique et la variance de u^*.
Nous avons pu distinguer trois formes notées $\mathcal{N}_1$, $\mathcal{N}_2$ et $\mathcal{N}_3$ de l'hypothèse d'absence de liaison, conduisant respectivement à une distribution hypergéométrique, binomiale ou de Poisson de la loi de probabilité de u^*. C'est une forme équivalente à $\mathcal{N}_2$ qui a été considérée dans (Gras 1979). Néanmoins, comme nous l'avons déjà exprimé, c'est $\mathcal{N}_1$ et $\mathcal{N}_3$ qui se distinguent le mieux (Lerman et al. 1981). En désignant par $q_i(a, \neg b)$, l'indice u centré et réduit par rapport à l'hypothèse $\mathcal{N}_i$, on a

$$q_1(a, \neg b) = q_1(\neg a, b) = -q_1(a, b) = -q_1(\neg a, \neg b) \quad (47)$$

Ainsi, pour l'hypothèse d'absence de liaison $\mathcal{N}_1$, la forme implicative de l'indice est exactement équivalente à la forme symétrique d'équivalence. Alors que pour la forme $\mathcal{N}_3$ de l'hypothèse d'absence de liaison, on a, avec $p(a) < p(b)$, on a

$$|q_3(a, \neg b)| > |q_3(b, \neg a)| \quad (48)$$

On peut considérer l'évaluation de l'implication $a \Rightarrow b$ dès lors que $n(a \wedge \neg b) - (\frac{n(a)*n(\neg b)}{n})$ est négatif. Cette quantité représente le numérateur de $q_3(a, \neg b)$. Elle est identique au numérateur $n(b \wedge \neg a) - (\frac{n(b)*n(\neg a)}{n})$ de l'indice $q_3(b, \neg a)$ qu'on aurait considéré relativement à l'implication inverse $b \Rightarrow a$. Cependant cette dernière est plus difficilement admissible puisque $n(b) > n(a)$. L'inégalité (48) est donc bienvenue, puisqu'en ce qui concerne l'indice local de la vraisemblance du lien, relatif à $\mathcal{N}_3$, on a (voir (45))

$$\mathcal{J}_3(a, b) > \mathcal{J}_3(b, a) \tag{49}$$

Pour s'en rendre compte, on peut se référer à l'excellente approximation normale - dès que n est assez grand (de l'ordre supérieur à quelques dizaines) - de la loi de Poisson de $n(a^* \wedge \neg b^*)$ [resp. $n(b^* \wedge \neg a^*)$] sous $\mathcal{N}_3$. Dans ces conditions

$$\mathcal{J}_3(a, b) = 1 - \Phi(q_3(a, \neg b)) \tag{50}$$

et

$$\mathcal{J}_3(b, a) = 1 - \Phi(q_3(b, \neg a)) \tag{51}$$

où Φ est la fonction de répartition de la loi normale centrée et réduite.

Même plus, la condition nécessaire et suffisante pour avoir (49) est $n(a) < n(b)$. En supposant comme c'est le cas le plus fréquent que $p(a) < p(b) < \frac{1}{2}$, on obtient la suite des inégalités suivantes, où celle (49) est comprise:

$$\mathcal{I}_3(\neg a, \neg b)) < \mathcal{J}_3(b, a)) < \mathcal{J}_3(a, b)) < \mathcal{I}_3(a, b) \tag{52}$$

où $\mathcal{I}_3$ est l'indice probabiliste local défini dans (18).
Nous allons maintenant considérer deux situations que nous noterons I et II où $\mathcal{O}(a)$ et $\mathcal{O}(b)$ gardent entre eux la même position relative , c'est-à-dire, où $n(a \wedge b)$, $n(a \wedge \neg b)$ et $n(\neg a \wedge b)$ gardent entre eux les mêmes rapports. Mais où, bien entendu, $n(\neg a \wedge \neg b)$ se comporte de façon relative différemment entre I et II. On remarque que tout indice d'équivalence, tel que par exemple celui de Jaccard (Jaccard 1908), se fondant sur des proportions relatives à l'intérieur de $\mathcal{O}(a \vee b)$, ne pourra distinguer entre I et II. Il en est également de l'indice dit confiance $(a \rightarrow b) = \frac{n(a \wedge b)}{n(a)}$. Il en sera tout autrement d'un indice d'implication dont la conception fera nécessairement intervenir $\mathcal{O}(\neg a \wedge \neg b)$.
On commence par supposer que $n = 4000$. La situation I est caractérisée par $n(a \wedge b) = 200$, $n(a \wedge \neg b) = 400$ et $n(\neg a \wedge b) = 600$; alors que la situation II concernant la comparaison d'un couple (a', b') d'attributs est caractérisée par $n(a' \wedge b') = 400$, $n(a' \wedge b') = 800$ et $n(\neg a' \wedge b') = 1200$. Ainsi, $n(a' \wedge b')$, $n(a' \wedge \neg b')$ et $n(\neg a' \wedge b')$ résultent de la multiplication par 2 de, respectivement $n(a \wedge b)$, $n(a \wedge \neg b)$ et $n(\neg a \wedge \neg b)$. Il ne faut pourtant pas s'étonner que la situation I dénote au sens de l'indice local une forte implication $a \Rightarrow b$ $(q_3(a, \neg b) = -3.65)$; alors que pour la situation II, l'implication

$a' \Rightarrow b'$ est inexistante ($q_3(a', \neg b') = 2.98$). En effet, compte tenu des degrés de liberté offerts par les tailles des ensembles mis en jeu, la pénétration de $\mathcal{O}(a)$ dans $\mathcal{O}(b)$ s'avère beaucoup plus exceptionnelle que celle de $\mathcal{O}(a')$ dans $\mathcal{O}(b')$. Cette dernière aurait dû atteindre la valeur $n(a' \wedge b') = 578$ pour avoir $q_3(a', \neg b') = q_3(a, \neg b) = -3.65$.
Le phénomène se trouve amplifié si on multiplie proportionnellement tous les cardinaus par un coefficient supérieur à 1. Ainsi avec un facteur 10 on obtient $q_3(a, \neg b) = -11.55$ et $q_3(a', \neg b') = 9.43$. Avec un facteur 100 on obtient $q_3(a, \neg b) = -36.51$ et $q_3(a', \neg b') = 29.81$.
Cette circonstance rend l'indice probabiliste local $\mathcal{J}_3(a, b)$ non discriminant dès lors qu'il y a lieu de comparer entre elles plusieurs implications. Pour y pallier on propose dans (Gras et al. 2001) de combiner au moyen d'une moyenne géométrique, d'une part l'indice $\mathcal{J}_3(a, b)$ (noté φ) et d'autre part, l'indice d'inclusion $\tau(a, b)$ [*cf.* (43)], pour aboutir à l'indice dit d'« intensité entropique » :

$$\Psi(a, b) = \sqrt{\varphi(a, b) * \tau(a, b)} \tag{53}$$

Mais alors, les indices $\varphi(a, b)$ et $\tau(a, b)$ correspondent à des philosophies bien distinctes. Alors que, sur les plans logique et statistique, ils sont liés d'une façon difficile à analyser; bien qu'un lien formel puisse être établi entre le χ^2 et la quantité d'information mutuelle associés à un tableau de contingence (Benzécri et Coll. 1973). De sorte que si n n'est pas trop grand pour permettre à $\varphi(a, b)$ d'être discriminant, on ne contrôle pas quelle va être « la part des choses » dans l'évaluation de l'indice $\Psi(a, b)$. D'autre part, pour n assez grand, $\varphi(a, b)$ va avoisiner 0 ou 1 ; de sorte que l'indice $\Psi(a, b)$ va se réduire à 0 ou à $\sqrt{\tau(a, b)}$. Les considérations formelles et statistiques (voir aussi la remarque suivant l'équation 45) peuvent avec intérêt être analysées sur le plan concret. Elles ne présument en rien l'intérêt des résultats qui ont pu être obtenus avec $\Psi(a, b)$.

4.3 Similarité implicative de vraisemblance du lien dans le contexte

La solution proposée ici et déjà exprimée dans (Lerman et al. 1981), mais sous une forme sensiblement moins élaborée et justifiée, pour arriver à un indice probabiliste d'implication discriminant quelle que soit la valeur de n, est celle de la réduction globale des similarités implicatives de la forme $q_3(a, \neg b)$. Pour un couple donné d'attributs (a, b), cette réduction doit s'effectuer par rapport à une base de couples d'attributs dont il y a lieu de comparer de façon mutuelle les implications. Le couple d'attributs (a, b) est l'un des éléments de cette base. Dans ces conditions, on se situe à l'intérieur d'une structure statistique qui peut être de forte liaison.
Cette méthode ne fait que transposer au cas dissymétrique la réduction globale des similarités symétriques [*cf.* partie 3] et qui nous a donné d'excellents résultats dans la pratique de la classification ascendante hiérarchique selon la vraisemblance du lien (méthode AVL) (Lerman 1991).
Un des problèmes posés consiste dans le choix de la base formée par les couples d'attributs qui va servir à la normalisation globale. Celui qui a été considéré dans l'analyse expérimentale qui suit a consisté à considérer tous les couples distincts d'attributs de $\mathcal{A}$ [*cf.* 19]. Nous l'indiquerons par son graphe

$$\mathcal{G}_0 = \{(j,k)|1 \leq j \neq k \leq p\} \tag{54}$$

Dans (Lerman et al. 1981) un choix sélectif a été préconisé où le graphe de référence est

$$\mathcal{G}_1 = \{(j,k)|(1 \leq j \neq k \leq p) \wedge (n(a^j) < n(a^k))\} \tag{55}$$

de sorte que l'absorption totale de $\mathcal{O}(a^j)$ par $\mathcal{O}(a^k)$ puisse se réaliser.
L'usage veut maintenant qu'on ne puisse avoir à considérer une implication de la forme $a \Rightarrow b$, que si des indices tels que $support(a \rightarrow b) = \frac{n(a \wedge b)}{n}$ et $confiance(a \rightarrow) = \frac{n(a \wedge b)}{n(a)}$ sont respectivement supérieurs à des seuils s_0 et c_0 que l'expert peut fixer (Agrawal et al. 1993, Guillaume 2000). Ainsi, nous préconisons la base indiquée par le graphe

$$\begin{aligned} \mathcal{G}_2 = \{(j,k)|(1 \leq j \neq k \leq p) \quad & \wedge \quad (n(a^j) < n(a^k)) \\ & \wedge \quad (support(a^j \rightarrow a^k) > s_0) \\ & \wedge \quad (confiance(a^j \rightarrow a^k) > c_0\} \end{aligned} \tag{56}$$

Relativement à un graphe $\mathcal{G}_i$ ($i = 0, 1$ ou 2), on substitue à l'indice « local » $q_3(a^j, a^k)$, celui « global » $q_3^g(a^j, a^k)$ qui se met sous la forme

$$q_3^g(a^j, a^k) = \frac{q_3(a^j, a^k) - moy_e\{q_3|\mathcal{G}_i\}}{\sqrt{var_e\{q_3|\mathcal{G}_i\}}} \tag{57}$$

où $moy_e\{q_3|\mathcal{G}_i\}$ et $var_e\{q_3|\mathcal{G}_i\}$ désignent respectivement la moyenne et la variance empirique de q_3 sur $\mathcal{G}_i$.
Le modèle de l'hypothèse d'absence de liaison est celui $\mathcal{N}_3$ déjà envisagé, mais où on associe à l'ensemble $\mathcal{A}$ des attributs observés, un ensemble $\mathcal{A}^*$ d'attributs aléatoires indépendants [*cf.* (28)]. Dans ces conditions, $q_3^g(a^{j*}, a^{k*})$ suit une loi normale centrée et réduite dont on indique par Φ la fonction de répartition. Ainsi, l'indice probabiliste discriminant (de vraisemblance du lien) d'une règle d'association ($a^j \Rightarrow a^k$) prenant place dans $\mathcal{G}_i$, s'exprime par

$$\mathcal{J}_n(a^j, a^k) = 1 - \Phi[q_3^g(a^j, a^k)] \tag{58}$$

5 Résultats expérimentaux

5.1 Le protocole expérimental

Nous avons réalisé deux séries d'expériences sur la base de données « mushrooms » (Blake et Merz 1998). Cette base est constituée de 22 attributs discrets, d'une classe et de 8124 individus. La classe a été assimilée à un simple attribut et l'ensemble des

attributs a été transformé en attributs booléens. Ainsi, après transformation, nous obtenons une base contenant 125 attributs qualificatifs booléens et 8124 individus.

Le choix de cette base est lié au « désir » d'observer le comportement des indices testés sur des données réelles plutôt qu'artificielles.

Le protocole expérimental utilisé est le suivant :

1. Isoler les couples (a, b) tel que $n(a) < n(b)$
2. Pour chacun de ces couples, calculer $q_3(a, \neg b)$ [*cf.* (7)] et $\mathcal{J}_3(a, b)$ [*cf.* (50)]
3. Ensuite, calculer la moyenne et l'écart-type de q_3 pour l'ensemble des couples (a, b) retenus et normaliser l'indice q_3 avant de faire appel à la loi normale.

Lors de la première série d'expériences, la sélection des couples (a, b) n'est conditionnée que par la contrainte suivante : $n(a) \neq 0$, $n(b) \neq 0$ et $n(a, b) \neq 0$.
Dans la deuxième série d'expériences, les contraintes appliquées aux couples sélectionnés sont liées au support et à la confiance de chacun des couples considéré. Le support d'un couple (a, b) est égal à $Pr(a \wedge b)$, sa confiance est égale à $Pr(b|a) = \frac{Pr(a \wedge b)}{Pr(a)}$. Ainsi, les couples (a, b) retenus, dans la deuxième série d'expériences, sont tels que $support(a, b) \geq seuil_{support}$ et $confiance(a, b) \geq seuil_{confiance}$.

L'objectif principal de cette seconde série d'expérienceces est d'observer le comportement des indices lorsque la taille de $\mathcal{O}$ augmente sans que les cardinaux de $\mathcal{O}(a)$, $\mathcal{O}(b)$ et $\mathcal{O}(a) \cup \mathcal{O}(b)$ soient modifiés. On retrouve ainsi des configurations statistiques de couples qu'on rencontre dans le cadre du « Data Mining ». Nous choisissons un couple (a, b) satisfaisant les contraintes du protocole et nous augmentons la valeur de n, sans modifier les valeurs $n(a)$, $n(b)$ et $n(a \wedge b)$. Tout se passe comme si on ajoutait des objets fictifs où tous les attributs sont à « faux ».

L'algorithme permettant de réaliser ces expériences est le suivant :

Algorithme 1

Entrée : E : ensemble des couples (a, b) satisfaisant les contraintes de l'expérience
Entrée : $seuil_n$: valeur maximale pour n
Tant que $(n < seuil_n)$ **Faire**
Pour tout $(a, b) \in E$ **Faire**
calcul de $q_3(a, \neg b)$
Fin Pour
Calcul de la moyenne et de l'écart-type des valeurs de l'indice q_3 obtenues
Pour tout $(a, b) \in E$ **Faire**
Calcul de l'intensité d'implication « classique » (II)
Calcul de l'intensité d'implication normalisée (IIN)
Fin Pour
$n \leftarrow n + constante$
Fin Tant que

Les résultats obtenus pour les deux séries d'expériences réalisées sont détaillés dans la partie suivante.

5.2 Les différents résultats

La première série d'expériences (aucune contrainte sur les couples (a, b)) n'a permis de mettre en évidence plusieurs propriétés intéressantes de l'indice normalisé).

- cet indice se montre discriminant quelle que soit la valeur de n
- de plus les résultats montrent que le comportement de l'indice est plus soutenu relativement aux implications $a \Rightarrow b$ lorsque $n(a) < n(b)$

Nous pouvons voir, sur les Figures 1 et 2, que l'indice normalisé présente un comportement discriminant, étant données les caractéristiques cardinales des attributs a et b mis en jeu, contrairement à l'indice classique (toujours égal à 1).

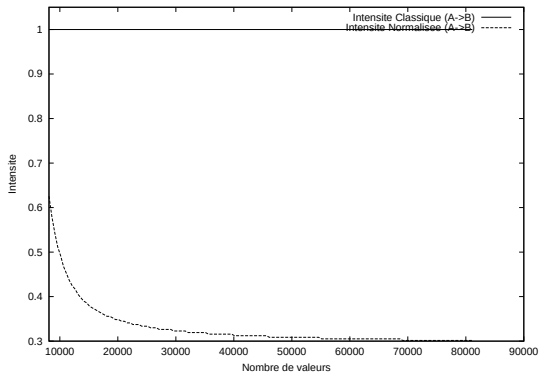

FIG. 1 – *Cas où* $n(a) = 192$, $n(b) = 1202$, $n(a \wedge b) = 96$, $s_0 = 0$, $c_0 = 0$

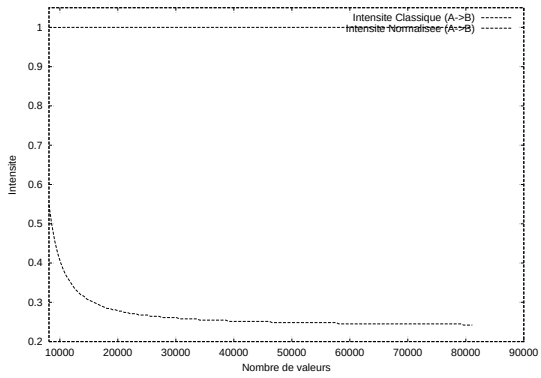

FIG. 2 – *Cas où* $n(a) = 192$, $n(b) = 828$, $n(a \wedge b) = 64$, $s_0 = 0$, $c_0 = 0$

Enfin, comme le montre la Figure 3, l'indice normalisé tend vers des valeurs relativement modérées, contrairement à l'indice classique (qui tend toujours vers 1), même lorsque la valeur de $n(a \wedge b)$ est relativement faible par rapport à $n(a)$ (voir la Figure 4 associée au résultat 3).

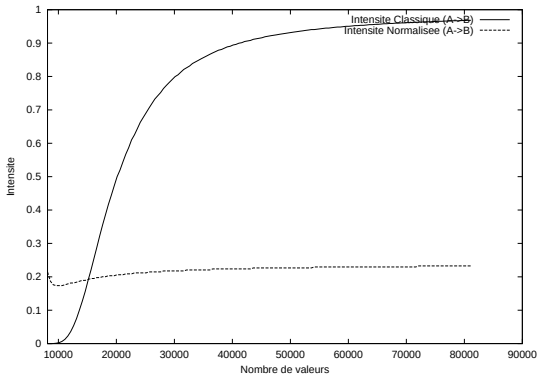

FIG. 3 – *Cas où* $n(a) = 452$, $n(b) = 2320$, $n(a \wedge b) = 52$, $s_0 = 0$, $c_0 = 0$

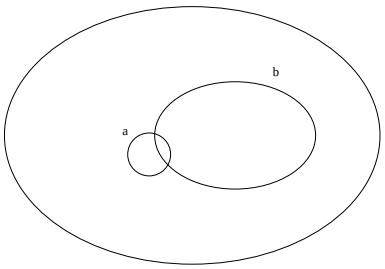

FIG. 4 – *Cas où* $n(a) = 452$, $n(b) = 2320$, $n(a \wedge b) = 52$

5.3 Robustification de l'indice d'implication normalisé (...on travaille sur un ensemble filtré de couples ...)

La deuxième série d'expériences permet de se focaliser sur des couples (a, b) ayant des propriétés de support et de confiance définies par l'utilisateur. Pour l'ensemble des expériences réalisées, nous avons fixé $s_0 = 0.1$ et $c_0 = 0.9$. Ainsi, nous ne retiendrons que les couples (a, b) tels que $support(a \rightarrow b) \geq 0.1$ et $confiance(a \rightarrow b) \geq 0.9$.

Ces seuils correspondent à des seuils relativement usuels en extraction de règles d'association et des valeurs similaires ont été souvent utilisées pour la base de données « mushrooms » (Lehn 2000, Bastide et al. 2002).

Les résultats obtenus montrent que sur cet ensemble réduit de couples (a, b), l'indice normalisé se montre toujours plus discriminant que l'indice classique. De plus, comme le montre la Figure 6, l'indice normalisé se montre plus discriminant dans ce cadre expérimental que dans la première série d'expériences (Figure 7). Ces courbes correspondent au cas présenté dans la Figure 5.

Dans la deuxième série d'expériences, les couples retenus correspondent à des relations « fortes » et la relation présentée sur la Figure 5 devient de moins en moins « intense » lorsque n augmente, comparativement aux autres relations présentes. Ainsi, l'opération

de filtrage (basée sur l'utilisation des seuils s_0 pour le support et c_0 pour la confiance) permet de retenir un ensemble de relations ayant des fortes valeurs des indices étudiés. L'association que nous étudions, Figure 5, devient donc de moins en moins pertinente, relativement aux autres relations respectant les conditions s_0 et c_0 imposées, lorsque la valeur de n augmente significativement.

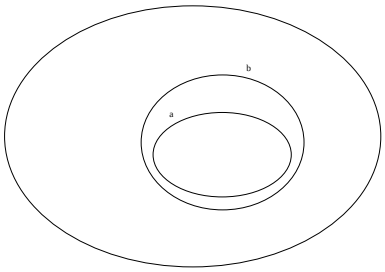

FIG. 5 – *cas où* $n(a) = 1296$, $n(b) = 2304$, $n(a \wedge b) = 1296$

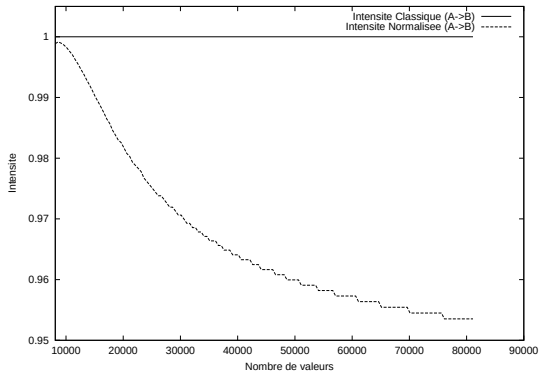

FIG. 6 – *cas où* $n(a) = 1296$, $n(b) = 2304$, $n(a \wedge b) = 1296$, $s_0 = 0$, $c_0 = 0$

Nous allons étudier relativement au protocole expérimental ci-dessus, la situation d'inclusion totale définie par $n(a) = 1296$, $n(b) = 2304$ et $n(a \wedge b) = 1296$, représentée par la Figure 5. La Figure 6 montre l'évolution de l'indice normalisé, calculé dans le contexte, lorsque le nombre d'objets augmente à partir de $n = 8124$ par adjonction d'objets où tous les attributs sont à « faux ». De toute façon, la valeur de l'indice se trouve discriminée. Elle demeure forte, supérieur à 0.98, pour n croissant jusqu'à 80000. Néanmoins, la valeur de l'indice va en décroissant vers un palier pour n augmentant. C'est que, pour n devenant gros, l'implication ci-dessus devient moins saisissante relativement à d'autres implications. Ces dernières peuvent ne pas correspondre à des inclusions totales ; mais, elles doivent concerner des situations où $n(a)$ et $n(b)$ sont sensiblement plus gros et proches.

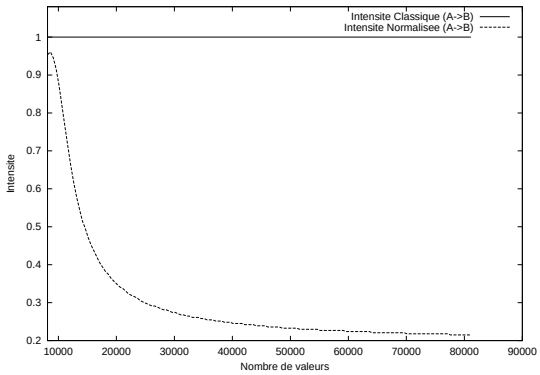

FIG. 7 – *cas où* $n(a) = 1296$, $n(b) = 2304$, $n(a \wedge b) = 1296$, $s_0 = 0.1$, $c_0 = 0.9$

Relativement à la dernière série d'expériences et à la Figure 7, il y a lieu d'ajouter qu'il ne faut pas s'étonner que la valeur de l'indice normalisé puisse tomber assez bas pour n augmentant. On se trouve en effet dans le contexte du graphe $\mathcal{G}_2$ [cf. (56)] des couples d'attributs (a_j, a_k) où la relation d'implication est très forte. Ainsi, en prenant $n = 80000$ (par l'adjonction d'objets, où tous les attributs sont à « faux »), on a pour chaque (a_j, a_k) retenu, $n(aj \wedge a_k) \geq 8000$ et $\frac{n(a_j \wedge a_k)}{n(a_j)} \geq 0.9$. Les indices se calculent sur la base des 8124 objets initiaux. Ils correspondent à des quasi-inclusions d'un très gros $\mathcal{O}(a_j)$ dans un à peine moins gros $\mathcal{O}(a_k)$.

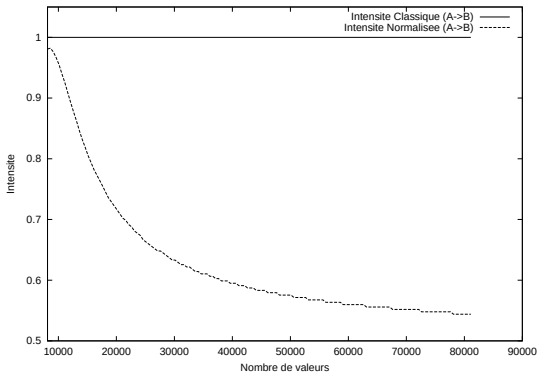

FIG. 8 – *cas où* $n(a) = 1296$, $n(b) = 2304$, $n(a \wedge b) = 1296$, $s_0 = 0.1$, $c_0 = 0.5$

Lorsque le seuil minimal de confiance est abaissé à 0.5, nous pouvons constater sur la Figure 8 que la relation étudiée devient plus significative par rapport à l'ensemble de relations considéré. Ce comportement est attendu et lié au fait que l'ensemble de relations étudié contient des relations beaucoup moins fortes que dans le cas de la Figure 7. Ce nouvel indice permet donc de mieux rendre compte du degré d'étonnement d'une

règle par rapport aux autres règles étudiées.

6 Conclusion

Un des objectifs fondamentaux de cet article est de montrer qu'il n'est pas nécessaire de sortir du cadre probabiliste de la vraisemblance du lien pour obtenir un indice discriminant dans l'évaluation quantifiée et deux à deux des règles d'association. Il importe pour cela de se placer de façon relative et de construire en conséquence un modèle aléatoire de l'hypothèse d'absence de liaison. L'accroissement de la complexité du calcul qui en résulte est linéaire par rapport à l'ensemble des règles d'intérêt en présence pour la normalisation.

Le travail présenté ici profite certes - comme nous l'avons mentionné - de résultats ou constructions préalablement obtenus, cependant, l'analyse menée ici, qui présente de multiples aspects nouveaux, est beaucoup plus systématique et focalisée, en se situant par rapport aux différentes approches exprimées dans le contexte de la fouille de données (« Data Mining »). Sur le chemin de la construction de l'indice probabiliste, nous voulons mettre l'accent sur l'indice $\eta_3(a, b)$ qui a un caractère local, ne dépendant que des seules proportions $p(a \wedge b)$, $p(a)$ et $p(b)$ et qui a été spécifié dans les conditions de cohérence $p(a) \leq p(b) \leq 0.5$.

Dans le cadre d'un même corpus de données, les indices probabilistes contextuels $\mathcal{J}_n(a^j, a^k)$ [cf. (58)] peuvent être obtenus à partir des indices $\eta_3(a^i, a^k)$ au moyen d'une transformation croissante.

Pourtant ce sont les valeurs numériques de $\mathcal{J}_n$ par rapport à celles de η_3 qui permettent de mieux mettre en avant et de mieux distinguer entre les différentes implications et ceci, par rapport au principe d'invraisemblance ou d'exceptionnalité qui existe bien dans la philosophie de la théorie de l'information. Maintenant, s'il s'agit de comparer une même règle $a \Rightarrow b$ dans les contextes différents de deux corpus, $\mathcal{J}_n$ le fait sur la même base relative. Enfin, il faut savoir que sur le plan mathématique, divers indices conçus dans un contexte local (cf 4.1) conduisent au même indice $\mathcal{J}_n$ par la construction que suppose la méthode de la vraisemblance de lien global. L'analyse expérimentale de validation menée est originale à plus d'un titre. Nous l'avons ici limitée à ce qui nous concernait au premier chef, à savoir, la comparaison entre l'indice probabiliste local et celui global pour l'intensité d'implication. Nous l'avons fait en nous appuyant sur la base de données classique des « mushrooms ». Bien que nous présumions que les résultats auraient été comparables dans le cadre d'autres bases de données, il aurait été intéressant de mener les expériences relativement à une autre base. Il y a lieu surtout d'étudier le comportement d'autres indices tels que $\psi(a, b)$ dit d'intensité entropique et celui $\eta_3(a, b)$ ci-dessus mentionné. Ces analyses expérimentales constitueront l'objet d'un prochain travail. Enfin, pour juger de l'intérêt d'une règle dans le contexte d'une base de données, en utilisant $\mathcal{J}_n$, des bornes peuvent être délimitées par l'expert lui-même en considérant les valeurs de $\mathcal{J}_n$ sur des situations d'inclusion telles que celle fournie dans la Figure 5.

Références

Agrawal, R., Imielinski, T., et Swami, A. N. (1993). Mining association rules between sets of items in large databases. Dans *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216.

Bastide, Y., Taouil, R., Pasquier, N., Stumme, G., et Lakhal, L. (2002). Pascal: un algorithme d'extraction des motifs fréquents. *Techniques et Science Informatiques*, 21(1):65–95.

Benzécri, J.P. et Coll., (1973). Théorie de l'information et classification d'après un tableau de contigence. Dans *L'Analyse des Données*, tome 1 : La Taxinomie, Dunod, Paris.

Bernard, J.-M. et Charron, C. (1996). L'analyse implicative bayésienne, une méthode pour l'étude des dépendances orientées: Données binaires. *Revue Mathématique Informatique et Sciences Humaines*, 134:5–38.

Blake, C. et Merz, C. (1998). UCI repository of machine learning databases.

Brin, S., Motwani, R., et Silverstein, C. (1997). Beyond market baskets: generalizing association rules to correlations. Dans *Proceedings of ACM SIGMOD'97*, pages 265–276.

Daudé, F. (1992). *Analyse et justification de la notion de ressemblance entre variables qualitatives dans l'optique de la classification hiérarchique par AVL*. PhD thesis, Université de Rennes 1.

Gras, R. (1979). *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques*. Master's thesis, Université de Rennes 1.

Gras, R., Kuntz, P., et Briand, H. (2001). Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille des données. *Revue Mathématique et Sciences Humaines*, 154-155:9–29.

Gras, R. et Larher, A. (1992). L'implication statistique, une nouvelle méthode d'analyse de données. *Mathématique Informatique et Sciences Humaines*, 120:5–31.

Guillaume, S. (2000). *Traitement des données volumineuses. Mesures et algorithmes d'extraction de règles d'association et règles ordinales*. PhD thesis, Université de Nantes.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise en Sciences Naturelles*, 44:223–270.

Lallich, S. et Teytaud, O. (2003). Evaluation et validation de l'intérêt des règles d'association. Soumission à GafoDonnées.

Lehn, R. (2000). *Un système interactif de visualisation et de fouille de règles pour l'extraction de connaissances dans les bases de données*. PhD thesis, Institut de Recherche en Informatique de Nantes.

Lerman, I., Gras, R., et Rostam, H. (1981). Elaboration et évaluation d'un indice d'implication pour des données binaires i. *Revue Mathématique et Sciences Humaines*, 75:5–35.

Lerman, I. C. (1984). Justification et validité statistique d'une échelle [0,1] de fréquence mathématique pour une structure de proximité sur un ensemble de variables observées. *Publications de l'Institut de Statistique des Universités de Paris*, 29:27–57.

Lerman, I. C. (1991). Foundations of the likelihood linkage analysis (lla) classification method. *Applied Stochastic Models and Data Analysis*, 7:63–76.

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs*, 61:1–49.

Pearson, K. (1900). On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling.

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. Dans *Knowledge Discovery in Databases*, pages 229 – 248. AAAI Press / The MIT Press.

Rodney M. Goodman, P. S. (1988). Information-theoretic rule induction. Dans *ECAI 1988*, pages 357–362.

Summary

The likelihood of the link probabilistic index, measuring an association rule, becomes no finely discriminant when the data size becomes large enough. The aim of this paper consists in showing the discriminant extension of this probabilistic index in order to measure an association rule in the context of a set of association rules. This method has been proposed for a long time and has been extensively validated in the framework of the AVL (Analyse de la Vraisemblance des Liens) hierachical clustering method of descriptive attributes. An experimental design is considered in order to establish the relevance of our statistical approach. This latter is also theoretically validated.

Keywords : Association rule, probabilistic discriminant index, validation.