Comparaison entre deux indices pour l'évaluation probabiliste discriminante des règles d'association

Israël-César Lerman*, Sylvie Guillaume**

*Irisa - Université de Rennes 1, Campus de Beaulieu, 35042 Rennes Cédex
lerman@irisa.fr,
**Clermont Université, Auvergne, LIMOS, BP 10448, F-63000 Clermont-Fd
guillaum@isima.fr

Résumé. L'élaboration d'une échelle de probabilité discriminante pour la comparaison mutuelle entre plusieurs attributs observés sur un échantillon d'objets de "grosse" taille, nécessite une normalisation préalable. L'objet de cet article est l'analyse comparée entre deux approches. La première dérive de l"'Analyse de la Vraisemblance des Liens Relationnels *Normalisée*". La seconde est fondée sur la notion de "Valeur Test" sur un échantillon *virtuel* de taille 100, synthétisant l'échantillon initial.

1 Introduction

Relativement à une base de données où on distingue un ensemble $\mathcal{A}$ d'attributs booléens observés sur un ensemble $\mathcal{O}$ d'entités (objets, individus, ...), le problème fondamental et bien connu en "Fouille des Données" ("Data Mining") est de pouvoir inférer un ensemble significatif et exploitable de règles d'association, on dit encore d'implications, entre attributs. Pour $(a, b) \in \mathcal{A} \times \mathcal{A}$, une règle de la forme $a \rightarrow b$ où a et b sont des attributs singletons, est un cas particulier d'une règle d'association. Intuitivement elle signifie que si a est à $VRAI$ sur un élément de $\mathcal{O}$, alors généralement - mais sans que cela soit un absolu - b est également à $VRAI$ sur cet élément de $\mathcal{O}$. Pour détecter de telles associations orientées, il importe de disposer d'un indice (on dit encore "coefficient" ou "mesure") statistique pertinent d'implication qui permette de dégager des "règles d'association" *intéressantes* ; c'est - à - dire, qui augmentent notre connaissance du réseau des tendances causales entre attributs de $\mathcal{A}$. Dans les travaux de Agrawal et al. (1993) qui ont lancé ces recherches dans le domaine de la "Fouille des Données" le souci d'une *bonne* mesure ne se manifestait pas encore. Il transparaît clairement dans Tan et al. (2002) où différents critères sont considérés pour organiser un ensemble de coefficients définissant des indices d'implication. D'autres études comparatives avec des facettes différentes sont également considérées dans Lallich et Teytaud (2004); Lenca et al. (2004). La mise à contribution de la notion d'indépendance statistique entre attributs intervient dans l'élaboration de nombreux coefficients Lallich et Teytaud (2004); Lenca et al. (2004); Lerman et Azé (2007); Piatetsky-Shapiro (1991); Tan et al. (2002). Historiquement, l'élaboration d'une échelle de probabilité pour éprouver l'existence d'un lien entre *deux* attributs descriptifs a été établie dans l'optique des tests d'hypothèses statistiques. L'adaptation au problème

de la comparaison mutuelle entre *plusieurs* attributs nécessite une normalisation préalable laquelle est indispensable pour que l'échelle de probabilité reste discriminante pour un nombre n d'observations augmentant de façon considérable (n pouvant atteindre plusieurs millions). C'est précisément ce caractère discriminant de l'échelle de probabilité qui permet d'obtenir un ensemble de règles de taille exploitable Azé (2003). L'objet de l'article est de comparer deux types d'indices. Le premier est issu de la méthodologie de l'*Analyse de la Vraisemblance des Liens Relationnels* Lerman (1970, 1973); Gras (1979); Lerman et al. (1981); Lerman (1981); Daudé (1992); Lerman et Azé (2007); Lerman (2009). Le second type d'indice correspond à une approche plus récente qui se réfère à l'optique des "tests statistiques d'hypothèses d'indépendance" Morineau et Rakotomalala (2006); Rakotomalala et Morineau (2008). En section 2, nous comparerons sur le plan conceptuel les deux approches VL ("Vraisemblance du Lien") et VT ("Valeur Test"). Nous présenterons en section 3 l'ensemble des indices normalisés ; l'*Intensité d'Implication Contextuelle* Lerman et Azé (2007) et les différentes versions de VTe où on propose une réduction synthétique de la taille n de l'échantillon initial à e. Deux de ces versions sont nouvelles. La comparaison théorique en section 4 des deux types d'indices normalisés se fera par rapport à deux modèles de croissance du nombre n d'observations notés M_1 et M_2. Nous conclurons en section 5 en donnant quelques perspectives.

2 Les approches VL et VT

En reprenant nos notations (voir section 1), relativement à un couple (a, b) d'attributs booléens de $\mathcal{A} \times \mathcal{A}$, nous introduisons les conjonctions $a \wedge b$, $a \wedge \bar{b}$, $\bar{a} \wedge b$ et $\bar{a} \wedge \bar{b}$ qui sont respectivement représentés par $O(a \wedge b) = O(a) \cap O(b)$, $O(a \wedge \bar{b}) = O(a) \cap O(\bar{b})$, $O(\bar{a} \wedge b) = O(\bar{a}) \cap O(b)$ et $O(\bar{a} \wedge \bar{b}) = O(\bar{a}) \cap O(\bar{b})$. Les cardinaux de ces sous ensembles sont respectivement désignés par $n(a \wedge b)$, $n(a \wedge \bar{b})$, $n(\bar{a} \wedge b)$ et $n(\bar{a} \wedge \bar{b})$. Ces cardinaux prennent place dans la table de contingence croisant les deux attributs binaires $\{a, \bar{a}\}$ et $\{b, \bar{b}\}$, voir le tableau 1. En rapportant les cardinaux à n, on définit les fréquences relatives ou proportions $p(a \wedge b)$, $p(a \wedge \bar{b})$, $p(\bar{a} \wedge b)$ et $p(\bar{a} \wedge \bar{b})$.

L'approche VL (*Vraisemblance du Lien*) Le cas symétrique de la comparaison entre deux attributs booléens a et b Lerman (1970, 1973, 1981); Daudé (1992); Lerman (2009) a précédé celui dissymétrique Gras (1979); Lerman et al. (1981); Lerman et Azé (2007) où on cherche à mettre en évidence la tendance implicative $a \rightarrow b$. Dans ce dernier cas on introduit un indice "brut" de la forme $n(a \wedge \bar{b})$ qui représente le nombre de contre exemples à la règle $a \rightarrow b$ et on cherche à évaluer la "petitesse" de ce nombre compte tenu des tailles $n(a)$ et $n(b)$. À cette fin, on introduit une hypothèse d'absence de liaison qui associe au couple d'attributs observé (a, b), un couple $(a^\star, b^\star)$ d'attributs aléatoires indépendants, de telle sorte que les espérances mathématiques de $card[\mathcal{O}(a^\star)]$ et de $card[\mathcal{O}(b^\star)]$ soient respectivement égales à $n(a)$ et $n(b)$. L'indice est alors fonction du degré d'*invraisemblance* de la petitesse relative de $n(a \wedge \bar{b})$. L'indice probabiliste de la vraisemblance du lien *orienté* de a vers b - qui a été appelé *Intensité d'Implication* - prend dans ces conditions la forme $\mathcal{I}(a \rightarrow b) = Pr\{n(a^\star \wedge \bar{b^\star}) > n(a \wedge \bar{b})\}$.

C'est le modèle de Poisson qui traduit de façon adéquate l'hypothèse d'absence de liaison Lerman (1973); Lerman et al. (1981). Pour ce modèle $n(a^\star \wedge \bar{b^\star})$ suit une loi de Poisson de paramètre $n(a) \times n(\bar{b})/n$. Ce dernier correspond à l'estimation de la moyenne de cette loi dans le cas qui nous concerne. Introduisons ici l'indice $n(a \wedge \bar{b})$ centré et réduit par rapport à

la moyenne et à l'écart type de l'indice aléatoire $n(a^\star \wedge \bar{b}^\star)$, lequel s'exprime par $Q(a,\bar{b}) = \{n(a \wedge \bar{b}) - [n(a) \times n(\bar{b})/n]\}/\sqrt{n(a) \times n(\bar{b})/n}$.

L'indice probabiliste $\mathcal{I}(a \rightarrow b)$ peut être calculé de façon exacte jusqu'à des valeurs très élévées de n en utilisant la loi de Poisson de paramètre $n(a) \times n(\bar{b})/n$. Cependant, l'excellente approximation de la loi de Poisson par la loi normale permet d'avoir le résultat de ce calcul au moyen de la formule analytique :

$$Pr\{n(a^\star \wedge \bar{b}^\star) > n(a \wedge \bar{b})\} = 1 - \Phi[Q(a,\bar{b})] = \Phi[-Q(a,\bar{b})] \quad (1)$$

où Φ désigne la fonction de répartition de la loi normale centrée et réduite. Compte tenu de la très grande précision de l'approximation pour n "grand", on identifiera $\mathcal{I}(a \rightarrow b)$ avec (1).

L'approche VT ($ValeurTest$) Pour l'approche VL l'hypothèse d'absence de liaison que nous notons $\mathcal{N}$, est nécessairement à rejeter Lerman (1984); Daudé (1992). Cependant, cette hypothèse revêt une importance cruciale pour déterminer une échelle de probabilité permettant la comparaison des liens entre attributs. Dans sa conception, l'approche VT reste accrochée à la théorie des tests d'indépendance. Ainsi, un rôle essentiel est dévolu au seuil critique ou $p - value$. Ce seuil se met ici sous la forme :

$$p = Pr^{\mathcal{N}}\{n(a^\star \wedge \bar{b}^\star) \leq n(a \wedge \bar{b})\} \quad (2)$$

La valeur test que nous noterons $VT(a \rightarrow b)$ est définie par "le nombre d'écarts types de la loi normale centrée et réduite qu'il faut dépasser" pour couvrir la probabilité complémentaire $1 - p$ Rakotomalala et Morineau (2008). On a donc :

$$VT(a \rightarrow b) = \Phi^{-1}(1-p) = \Phi^{-1}[\mathcal{I}(a \rightarrow b)] \quad (3)$$

Compte tenu de l'identification entre $\mathcal{I}(a \rightarrow b)$ et (1), on a : $VT(a \rightarrow b) = -Q(a,\bar{b})$. Un indice tel que $Q(a,\bar{b})$ se présente sous la forme d'un coefficient de corrélation que multiplie $\sqrt{n}$ Lerman et al. (1981); Lerman et Azé (2007). Il s'ensuit que dans le cadre d'un modèle de croissance tel que M_1 (voir section 4), l'indice p (resp., $\mathcal{I}(a \rightarrow b)$) tend "très rapidement" vers 0 ou 1 (resp., 1 ou 0) selon que $n(a \wedge b) > n(a) \times n(b)/n$ ou que $n(a \wedge b) < n(a) \times n(b)/n$. Ainsi, l'échelle fondée sur VL ou VT ne permet plus de discriminer les liaisons entre différents attributs booléens observés sur le même ensemble d'objets dès lors que ce dernier a une taille élevée.

3 L'indice VL normalisé et trois versions de VTe

VL normalisée *VLgrImpP* Le principe de la réduction globale des indices d'association a déjà été proposé dans Lerman et al. (1981), il a été repris et étudié de façon consistante dans Lerman et Azé (2007). Relativement à une base de données considérons un ensemble d'attributs booléens que nous notons $\mathcal{A} = \{a^j \mid 1 \leq j \leq m\}$ sur lequel il y a lieu d'élaborer un indice d'implication. Nous allons distinguer dans l'ensemble $\mathcal{A} \times \mathcal{A}$ des couples d'attributs, un sous ensemble potentiel $\mathcal{C}$ de couples d'attributs (a,b) pour lesquels l'implication $a \rightarrow b$ "peut avoir un sens". Une première condition qu'on peut exiger pour un couple d'attributs (a,b) entrant dans la composition de $\mathcal{C}$ est d'être tel que $n(a \wedge \bar{b}) < n(a) \times n(\bar{b})/n$.

Cette condition est d'ailleurs équivalente à $n(a \wedge b) > n(a) \times n(b)/n$. Une deuxième condition qu'on peut exiger pour qu'un couple d'attributs (a, b) rentre dans la composition de $\mathcal{C}$, relativement à l'implication $a \rightarrow b$, est $n(a) < n(b)$. En effet, dans le cas où l'implication $a \rightarrow b$ est observée exactement (sans contre exemple), l'ensemble $\mathcal{O}(a)$ des individus où a est à $VRAI$ est strictement inclus dans l'ensemble $\mathcal{O}(b)$ des individus où b est à $VRAI$. D'ailleurs, relativement à un couple (a, b) d'attributs pour lequel $n(a) < n(b)$, on démontre que $Q(a, \bar{b}) < Q(b, \bar{a})$. La première version du sous ensemble potentiel $\mathcal{C}$ de couples d'attributs par rapport auquel la réduction globale peut être opérée est définie par $\mathcal{C}_0 = \Big\{(a, b) \mid (a, b) \in \mathcal{A} \times \mathcal{A}, n(a \wedge b) > n(a) \times n(b)/n$ et $n(a) < n(b)\Big\}$.

Dans ce cas, la réduction globale est dite *totale*. Nous nous limiterons dans cet article à cette dernière. Considérons dans ces conditions la distribution de $Q(a, \bar{b})$ sur $\mathcal{C}_0$ et désignons par $moy_0(Q)$ et $var_0(Q)$ la moyenne et la variance de cette distribution. Dans ces conditions, la distribution de l'indice globalement réduit sur $\mathcal{C}_0$,
$\{Q^{g0}(a, \bar{b}) = [Q(a, \bar{b}) - moy_0(Q)]/\sqrt{var_0(Q)} \mid (a, b) \in \mathcal{C}_0\}$ est de moyenne nulle et de variance unité. De sorte que la distribution des indices probabilistes de l'intensité d'implication

$$\{\mathcal{I}^0(a \rightarrow b) = \Phi(-Q^{g0}(a, \bar{b})) \mid (a, b) \in \mathcal{C}_0\} \tag{4}$$

devient finement discriminante pour comparer entre elles les différentes implications $a \rightarrow b$ où $(a, b) \in \mathcal{C}_0$.

L'approche VTe fondée sur la moyenne des p-values *VTeImpBarP* L'idée proposée dans Morineau et Rakotomalala (2006); Rakotomalala et Morineau (2008) est de se ramener à un échatillon *virtuel* de taille $e = 100$ synthétisant l'échantillon initial de taille n. Pour cette réduction *la solution* considérée par les auteurs consiste d'abord à substituer au tableau de contingence observé où n est "grand", voir le tableau 1, celui ramené à 100, voir le tableau 2.

	a	$\bar{a}$	Total
b	$n(a \wedge b)$	$n(\bar{a} \wedge b)$	$n(b)$
$\bar{b}$	$n(a \wedge \bar{b})$	$n(\bar{a} \wedge \bar{b})$	$n(\bar{b})$
Total	$n(a)$	$n(\bar{a})$	n

TAB. 1 – *Tableau de contingence* 2×2.

	a	$\bar{a}$	Total
b	$100 \times p(a \wedge b)$	$100 \times p(\bar{a} \wedge b)$	$100 \times p(b)$
$\bar{b}$	$100 \times p(a \wedge \bar{b})$	$100 \times p(\bar{a} \wedge \bar{b})$	$100 \times p(\bar{b})$
Total	$100 \times p(a)$	$100 \times p(\bar{a})$	100

TAB. 2 – *Tableau de contingence* 2×2 *réduit à* 100.

Cependant, dans ce dernier tableau où le total est ramené à 100, les contenus des cases ne sont plus des entiers, mais des rationnels dont on donne approximation décimale. On considère

alors le contenu de la case $(a, \bar{b})$ de ce dernier tableau 2 ainsi que les contenus des cases marginales qui l'encadrent. Soit donc, $\gamma = 100 \times p(a \wedge \bar{b})$, $\alpha = 100 \times p(a)$ et $\bar{\beta} = 100 \times p(\bar{b})$. On considère alors les 8 vecteurs à composantes entières les plus proches du vecteur $(\alpha, \bar{\beta}, \gamma)$. Un principe de moyenne barycentrique permet de retrouver le vecteur $(\alpha, \bar{\beta}, \gamma)$ à partir de ces 8 vecteurs. À chacun d'entre eux on associe la *p-value* pour le contenu de la case $(a, \bar{b})$ et on détermine la valeur moyenne de ces *p-values* ; cette dernière moyenne étant pondérée conformément aux coefficients de la moyenne barycentrique mentionnée. Cette *p-value* moyenne donne alors lieu à $VT100BarP$.

L'approche VTe fondée sur une approche corrélative et ensembliste *VTeImpCorP* Désignons ici l'ensemble $\mathcal{O}$ des objets sous la forme $\{o_i \mid 1 \leq i \leq n\}$ et introduisons les fonctions indicatrices des sous ensembles $\mathcal{O}(a)$ et $\mathcal{O}(b)$ (voir section 2) que nous notons également sans risque de confusion, a et b : $a(i) = 1$ (resp. 0) si et seulement si l'attribut a est à $VRAI$ (resp. $FAUX$) sur l'objet o_i, $1 \leq i \leq n$; de même, $b(i) = 1$ (resp. 0) si et seulement si l'attribut b est à $VRAI$ (resp. $FAUX$) sur l'objet o_i, $1 \leq i \leq n$. Dans ces conditions, l'indice "brut" $n(a \wedge \bar{b})$ et celui aléatoire associé $n(a^\star \wedge \bar{b^\star})$ s'écrivent respectivement :

$$n(a \wedge \bar{b}) = \sum_{i=1}^{n} a(i) \times [1 - b(i)] \text{ et } n(a^\star \wedge \bar{b^\star}) = \sum_{i=1}^{i=n} a[\sigma(i)] \times (1 - b[\tau(i)]) \qquad (5)$$

où σ et τ sont deux permutations aléatoires indépendantes prises dans un ensemble G_n de permutations sur $\{1, ..., i, ..., n\}$, muni d'une probabilité dépendant du modèle de l'hypothèse d'absence de liaison. Initialement, cette forme permutationnelle a été considérée pour la comparaison d'attributs numériques où un modèle de Poisson de l'hypothèse d'absence de liaison a été précisé. Pour notre problème et pratiquement, on se base sur le tableau 2 ci-dessus et on se réfère à un modèle de l'hypothèse d'absence de liaison de Poisson ayant un caractère permutationnel, relativement à l'indice "brut", $100 \times \big(n(a \wedge \bar{b})/n\big) = 100 \times p(a \wedge \bar{b})$ qui occupe la case $(a, \bar{b})$ de ce tableau. À cette fin et conceptuellement, nous allons procéder à la construction d'un ensemble virtuel Ω formé de 100 éléments : $\Omega = \{\omega_j \mid 1 \leq j \leq 100\}$. Sans que ces éléments aient une définition explicite, nous définissons sur ces derniers de façon potentielle, deux attributs α et β à valeurs dans l'intervalle $[0, 1]$, tels que :

$$\sum_{1 \leq j \leq 100} \alpha(j) \times (1 - \beta(j)) = 100 \times \big(n(a \wedge \bar{b})/n\big)$$
$$\sum_{1 \leq j \leq 100} \alpha(j) = 100 \times \big(n(a)/n\big) \quad \sum_{1 \leq j \leq 100} \beta(j) = 100 \times \big(n(b)/n\big) \qquad (6)$$

Les moyennes des variables α et β sont respectivement égales à $n(a)/n$ et $n(b)/n$. Cette construction peut être effectuée de différentes façons. Celle choisie est telle que soient maximales les variances de α et de β ; car ainsi, $var(\alpha) = var(a)$ et $var(\beta) = var(b)$. De la sorte, on démontre que $Q(\alpha, \bar{\beta}) = \sqrt{\frac{99}{n-1}} \times Q(a, \bar{b})$ Lerman et Guillaume (2010) et l'indice probabiliste de la *vraisemblance du lien* traduisant l'intensité de l'implication devient :

$$\mathcal{I}(a \rightarrow b) = Pr\{n(\alpha^\star \wedge \bar{\beta^\star}) > n(\alpha \wedge \bar{\beta})\} = \Phi(-Q(\alpha, \bar{\beta})) \qquad (7)$$

L'approche VTe obtenue par projection sur un ensemble aléatoire *VTeImpProj* Tel qu'il est présenté intuitivement dans Morineau et Rakotomalala (2006); Rakotomalala et Morineau (2008) l'indice VTe ($e = 100$ pour les auteurs), est sur le plan conceptuel complètement déconnecté de celui censé l'approximer et qui est $VTeImpBarP$. En effet, dans la présentation donnée, il s'agit de proposer la *moyenne* de l'indice $-Q(a, \bar{b})$ calculée sur une suite $(E^{(1)}, E^{(2)}, ..., E^{(l)}, ..., E^{(L)})$ de L échantillons aléatoires indépendants (on suppose que le tirage est sans remise) de taille e chacun. Indépendamment du problème de la détermination de L, une telle procédure, comme d'ailleurs c'est admis dans Morineau et Rakotomalala (2006), est très lourde. En fait elle est *inutile* puisqu'un calcul mathématique la remplace et permet de proposer une solution simple et efficace. L'expression mathématique de l'indice qui correspond à l'indice visé est :

$$\mathcal{E}\Big(\frac{card[\mathcal{O}(a) \cap \mathcal{O}(\bar{b}) \cap E^\star] - \frac{card[\mathcal{O}(a) \cap E^\star \times card[\mathcal{O}(\bar{b}) \cap E^\star]]}{e}}{\sqrt{\frac{card[\mathcal{O}(a) \cap E^\star \times card[\mathcal{O}(\bar{b}) \cap E^\star]]}{e}}}\Big) \tag{8}$$

où $E^\star$ est un sous ensemble aléatoire de taille e pris dans l'ensemble des parties de $\mathcal{O}$ de même cardinal e, muni d'une probabilité uniformément répartie et où $\mathcal{E}$ désigne l'espérance mathématique. La solution adoptée pour des raisons de complexité analytique, est :

$$\frac{\mathcal{E}\Big(card[\mathcal{O}(a) \cap \mathcal{O}(\bar{b}) \cap E^\star] - \frac{card[\mathcal{O}(a) \cap E^\star \times card[\mathcal{O}(\bar{b}) \cap E^\star]]}{e}\Big)}{\sqrt{\mathcal{E}\Big(\frac{card[\mathcal{O}(a) \cap E^\star \times card[\mathcal{O}(\bar{b}) \cap E^\star]]}{e}\Big)}} \tag{9}$$

On démontre que le calcul pour cette expression donne Lerman et Guillaume (2010) :

$$\frac{\frac{1}{\sqrt{n(n-1)}} \times [(ne-1) \times n(a \wedge \bar{b}) - (e-1) \times n(a) \times n(\bar{b})]}{\sqrt{(n-e) \times n(a \wedge \bar{b}) + (e-1) \times n(a) \times n(\bar{b})}} \tag{10}$$

La *Valeur Test* est l'opposée de cette dernière valeur.

4 Variations de $VLgrImpP$ et $VTeImpCorP$

Les deux modèles de croissance M_1 et M_2 Le modèle M_1 est classique. Pour un couple d'attributs booléens (a, b), les cardinaux n , $n(a)$, $n(\bar{a})$, $n(b)$, $n(\bar{b})$, $n(a \wedge b)$, $n(a \wedge \bar{b})$, $n(\bar{a} \wedge b)$ et $n(\bar{a} \wedge \bar{b})$, $n(a \wedge b)$ sont multipliés par une constante k. La valeur de k va croître dans le cadre du modèle. C'est ce modèle qui est considéré dans Morineau et Rakotomalala (2006); Rakotomalala et Morineau (2008). Pour le second modèle M_2 qui est spécifique, $n(a \wedge b)$, $n(a)$ et $n(b)$ sont invariables, seul n augmente à partir de sa valeur initiale considérée dans une situation réelle. Ainsi, on accroît le nombre $n(\bar{a} \wedge \bar{b})$ d'éléments pour lesquels a et b sont à $FAUX$. Si x définit cet accroissement, la suite des cardinaux précédents devient $n + x$, $n(a)$, $n(\bar{a}) + x$, $n(b)$, $n(\bar{b}) + x$, $n(a \wedge b)$, $n(a \wedge \bar{b})$, $n(\bar{a} \wedge b)$ et $n(\bar{a} \wedge \bar{b}) + x$. Ce modèle M_2 nous avait été suggéré par Y. Kodratoff (communication personnelle analysée dans Lerman et Azé (2007)). Il se justifie de façon tout à fait pertinente compte tenu de la taille des bases de données actuelles et compte tenu du fait que pour un attribut booléen donné décrivant un aspect de la base, le nombre d'entités où cet attribut est à $VRAI$ est en général "très petit"

par rapport à la taille de la base. Relativement au modèle M_2 nous allons comparer sur le plan mathématique les variations de l'indice $VLgrImpP$ d'une part et de l'indice $VTeImpCorP$ d'autre part. En effet ce dernier indice est pris comme représentant des indices VTe et c'est lui qui se prête le mieux à l'étude analytique. C'est l'analyse expérimentale évoquée en section 5 qui permettra une vision synthétique des comportements relatifs des quatre différents indices.

Comportement de $VLgrImpP$ et $VTeImpCorP$ par rapport au modèle M_1 Si ν désigne le nombre total d'observations du tableau de contingence obtenu à partir du tableau de contingence initial - croisant $\{a, \bar{a}\}$ avec $\{b, \bar{b}\}$ - par application du modèle M_1 ou M_2, le nouveau tableau pour lequel l'indice $Q(a, \bar{b})$ est calculé, s'obtient en multipliant les contenus des différentes cases du tableau par le rapport $100/\nu$. On a les deux résultats suivants.

Proposition 1 *L'indice $VLgrImpP$ est invariant pour n croissant conformément au modèle M_1.*

Proposition 2 *L'indice $VTeImpCorP$ est invariant pour n croissant conformément au modèle M_1.*

Comportement par rapport au modèle M_2 de $VLgrImpP$ Nous venons de voir ci-dessus ce que deviennent les cardinaux de la table de contingence croisant $\{a, \bar{a}\}$ et $\{b, \bar{b}\}$. L'indice $Q(a, \bar{b})$ associé que nous notons maintenant $Q_x(a, \bar{b})$ s'écrit :

$$Q_x(a, \bar{b}) = \frac{n(a \wedge \bar{b}) - \frac{n(a) \times [n(\bar{b})+x]}{(n+x)}}{\sqrt{\frac{n(a) \times [n(\bar{b})+x]}{(n+x)}}} \tag{11}$$

Nous considérerons plus précisément le coefficient $-Q_x(a, \bar{b})$ qui est une fonction croissante de l'indice probabiliste local d'implication de la *vraisemblance du lien* dit d"'*Intensité d'Implication*". L'étude du sens de variation par rapport à x de ce coefficient permet d'établir le résultat suivant :

Proposition 3 *$-Q_x(a, \bar{b})$ est croissant par rapport à x, son taux de variation est décroissant par rapport à x.*

Donnons ici une suite de valeurs de $-Q_x(a, \bar{b})$ dans le cas d' un exemple issu de la base de données "Adult" disponible sur le site "UCI Machine Learning Repository" et déjà utilisée dans (Morineau et Rakotomalala, 2006, voir section 5), où : $n = 14743$, $n(a) = 4819$, $n(\bar{b}) = 3522$ et $n(a \wedge \bar{b}) = 225$.

x	$-Q_x(a, \bar{b})$
0	27.712
1000	31.599
2000	34.687
10000	47.503
50000	60.234
100000	63.233

Considérons l'indice normalisé $Q^{g0}(a,\bar{b})$ et celui associé $\Phi[-Q_x^{g0}(a,\bar{b})]$. Si $-Q_x(a,\bar{b})$ est croissant par rapport à x, sa version normalisée $-Q_x^{g0}(a,\bar{b})$ ne l'est pas nécessairement toujours. Cette dernière s'avère néanmoins dans la pratique globalement monotone et cela de façon cohérente compte tenu de la situation de départ.

Comportement par rapport au modèle M_2 de $VT100ImpCorP$ Nous indiquons par $-Q_x^{100}$ la version adoptée de l'indice $VT100$. On obtient :

$$-Q_x^{100}(a,\bar{b}) = -10 \times \frac{[n(a \wedge \bar{b}) - \frac{n(a)\times[n(\bar{b})+x]}{n+x}]}{\sqrt{n(a) \times [n(\bar{b})+x]}} \tag{12}$$

Pour simplifier nos notations nous posons : $\gamma = n(a \wedge \bar{b})$, $\alpha = n(a)$, $\beta = n(b)$, $\bar{\beta} = n(\bar{b})$ et $y = n(\bar{b}) + x$. L'étude du sens de variation de cet indice lorsque x varie (à partir de 0 en croissant) conduit à l'étude du signe d'un trinôme du second degré en y dont le discriminant est $\Delta = \beta^2 \times [(\alpha+2\gamma)^2 + 4(\alpha-\gamma)]$. Les deux racines s'écrivent $y' = [\beta(\alpha+2\gamma)+\sqrt{\Delta}]/2(\alpha-\gamma)$ et $y'' = [\beta(\alpha+2\gamma) - \sqrt{\Delta}]/2(\alpha-\gamma)$. On obtient le résultat suivant.

Proposition 4 *L'indice $-Q_x^{100}(a,\bar{b})$ est croissant pour x variant dans l'intervalle $[0, y'-n(\bar{b})]$ et décroissant pour x supérieur à $y' - n(b)$.*

À titre d'illustration considérons l'exemple ci-dessus de la base de données "Adult". On obtient :

x	$-Q_x^{100}(a,\bar{b})$
0	2.282
1000	2.518
2000	2.681
10000	3.020
15000	2.973
20000	2.887
30000	2.694
50000	2.367
100000	1.982

Un tel comportement de $-Q_x^{100}(a,\bar{b})$, d'abord croissant puis décroissant peut surprendre. Il se trouve confirmé par l'analyse expérimentale.

5 Conclusion et perspectives

Cette étude est essentiellement théorique. Elle a été appuyée par une analyse expérimentale qui a utilisé la base de données "Wages" Franck et Asuncion (2010). Cette analyse a parfaitement validé les résultats théoriques établis (Lerman et Guillaume, 2010, voir section 6). C'est ainsi que l'invariance de chacun des quatre indices a été constatée pour le modèle d'évolution M_1 (voir Propositions 1 et 2). Toutefois, l'analyse expérimentale révèle des aspects que l'analyse théorique ne peut sonder. Un plan d'expérience organisé selon le degré de dépendance

implicative (Lerman et Guillaume, 2010, voir section 6) permet de déceler des différences de comportement dans l'évolution des quatre indices les uns par rapport aux autres. À cet égard, l'indice $VLgrImpP$ apparaît comme devant jouer un rôle de référence. Ainsi pour le modèle M_2, il s'agit du seul indice à avoir un sens de variation monotone ; soit croissant soit décroissant, selon la structure de la liaison implicative. La croissance ou la décroissance se fait d'abord avec une forte pente qui s'adoucit petit à petit pour devenir de plus en plus proche de l'horizontale. Maintenant, une question fondamentale concerne le choix des règles les plus fortement quantifiées pour l'un ou l'autre des quatre indices. C'est le caractère discriminant de l'indice qui permet ce choix. Il sera à étudier pour l'un ou l'autre des indices dans le cadre du modèle M_2. Ainsi, on peut limiter *a priori* le nombre de règles extraites ; mais sans savoir nécessairement le nombre de règles qu'il y a lieu d'extraire. En cas d'indices discriminants, est-ce que les m règles ($m = 1, 2, ...$) les plus fortement quantifiées pour l'un ou l'autre des indices sont nécessairement identiques ? Sinon, peut - on caractériser les types de règles extraites par l'un ou l'autre des indices ? L'ensemble de ces questions, cruciales pour l'utilisateur sera abordé dans l'article suivant "*Analyse du comportement limite d'indices probabilistes pour une sélection discriminante*".

Références

Agrawal, R., T. Imielsky, et A. Swami (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 8th ACM SIGKDD Conference on Knowledge Discovery and Data Mining 1993*, ACM, pp. 207–216.

Azé, J. (2003). *Extraction de connaissances à partir de données numériques et textuelles*. Thèse de doctorat, Université de Paris Sud.

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

Franck, A. et A. Asuncion (2010). UCI Machine Learning Repository. Technical report, School of Information and Computer Science, University of California, Irvine.

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

Lallich, S. et O. Teytaud (2004). Évaluation et validation de l'intérêt des règles d'association. In *Mesures de Qualité pour la Fouille des Données 2004*, Volume RNTI-E-1 of *RNTI*, pp. 193–218. Cépaduès.

Lenca, P., P. Meyer, B. Picouet, et S. Lallich (2004). Évaluation et analyse multicritère des mesures de qualité des règles d'association. In *Mesures de Qualité pour la Fouille des Données 2004*, Volume RNTI-E-1 of *RNTI*, pp. 219–246. Cépaduès.

Lerman, I.-C. (1970). Sur l'analyse des données préalable à une classification automatique ; proposition d'une nouvelle mesure de similarité. *Mathématiques et Sciences Humaines 8*, 5–15.

Lerman, I.-C. (1973). Introduction à une méthode de classification automatique illustrée par la recherche d'une typologie des personnages enfants à travers la littérature enfantine. *Revue*

de Statistique Appliquée XXI, 23–49.

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

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. (2009). Analyse de la vraisemblance des liens relationnels : une méthodologie d'analyse classificatoire des données. In Y. Bennani et E. Viennet (Eds.), *Apprentissage artificiel et fouille de données 2009*, Volume RNTI A3 of *RNTI*, pp. 93–126. Cépaduès.

Lerman, I.-C. et J. Azé (2007). A new probabilistic measure of interestingness for association rules, based on the likelihood of the link. In *Quality measures in data mining 2007*, Volume 43 of *Studies in Computational Intelligence*, pp. 207–236. Springer.

Lerman, I.-C., R. Gras, et H. Rostam (1981). Élaboration et évaluation d'un indice d'implication pour des données binaires i et ii. *Mathématiques et Sciences Humaines 74-75*, 5–35, 5–47.

Lerman, I.-C. et S. Guillaume (2010). Analyse comparative d'indices d'implication discriminants fondés sur une échelle de probabilité. Rapport de recherche, INRIA, Rennes. 7187, Février 2010, 85 pages.

Morineau, A. et R. Rakotomalala (2006). Critère VT100 de sélection des régles d'association. In G. Ritschard et C. Djeraba (Eds.), *Actes de Extraction et Gestion de Connaissances, 2006*, Volume EGC'2006 of *RNTI*, pp. 581–592. Cépaduès.

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. In *Knowledge Discovery in Databases 1991*, pp. 229–248. MIT Press.

Rakotomalala, R. et A. Morineau (2008). The TVpercent principle for the counterexamples statistic. In R. Gras, E. Suzuki, F. Guillet, et F. Spagnolo (Eds.), *Statistical Implicative Analysis, 2008*, pp. 449–462. Springer.

Tan, P.-N., V. Kumar, et J. Srivastava (2002). Selecting the right interestingness measure for association patterns. In *Proceedings of the 8th ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2002*, ACM, pp. 32–41.

Summary

Preliminary standardization is needed for probabilistic pairwise comparison between descriptive attributes in Data Mining. The goal of this paper consists of comparison between two approaches. The first one is due to a normalized version of the "Likelihood Linkage Analysis" approach. The second one is based on the notion of "Test Value" defined with respect to a hypothetical sample, sized 100 and summarizing the initial observed sample.