Notes sur les mesures probabilistes de la qualité des règles d'association : un algorithme efficace d'extraction des règles d'association implicative

André Totohasina[⊗], Henri Ralambondrainy[*], Jean Diatta[*]

⊗ *Département de Mathématiques et Informatique, Ecole Normale Supérieure pour l'Enseignement Technique (ENSET) Université d'Antsiranana -B.P. O – Antsiranana, 201 – Madagascar* totohasina@yahoo.fr

*Institut de Recherche en Mathématiques et Informatique et leurs Applications (IREMIA) Université de La Réunion – 15, Avenue René Cassin- B.P. 751.97715, Saint-Denis, Messag Cedex 9, France. **{jean.diatta,** ralambon}@univ-reunion.fr

RÉSUMÉ. A la lumière de très riches propriétés de la mesure implicative orientée nommée centrée, non symétrique, définie dans (17), cet article propose un cadre unificateur des critères de qualité des règles d'association booléenne, via des mesures probabilistes, et un algorithme permettant de sélectionner des règles d'association interprétables en terme d'implication statistique, avec prise en compte de la surprise et de la cohérence avec une dépendance orientée positivement ou négativement. Son application en classification et la possibilité de son extension au traitement des variables latticielles (i.e. variables qualitatives hiérarchiques ou intervalles) seront présentées avant de conclure.

MOTS-CLÉS : Extraction des connaissances dans des données, règles d'association, qualité, probabilité conditionnelle, implication, mesures normalisées, algorithme, dépendance orientée.

ABSTRACT With the light of the very rich properties of the implicative measure as defined in (17), this paper proposes a unifying framework about probabilistic measures of interestingness of Boolean association rules, and an algorithm of mining association rules which are interpretable in term of statistical implication taking account of positively or negatively oriented dependency. Its possible applying in the clustering and extension for dealing with latticial variables (i.e. categorical, hierarchical or interval variables) will be developed before discussion and conclusion.

KEY-WORDS : Data mining; association rules, interestingness, conditional probability, normalized and centered measure, implication, normalisation of measures, algorithm; oriented correlation.

1. Préliminaire

Grossièrement, le problème de fouille des règles d'association qualitative comprend deux sous-problèmes [13] : Trouver tous les motifs fréquents, et Générer les règles d'association dérivant des motifs fréquents. Cependant l'ensemble des règles ainsi générées sont généralement de très grande taille. Aussi, se pose les problèmes de trouver des représentations compactes et de trouver des critères de qualité plus pertinents et sélectifs permettant d'éviter cette surabondance de règles extraites.

2. Introduction

Dans ce texte, nous considérons un espace probabilisé discret fini $(\Omega, \mathcal{T}(\Omega), P)$ tel que pour tout événement E de $\mathcal{T}(\Omega)$, Card(E) = cardinal(E), P(E)= Card(E)/Card(Ω).

Notations et terminologies : Ω : l'ensemble des n individus ou objets, sur lesquels on a mesuré m variables binaires $a_1, \ldots, a_m$, $\Gamma = \{a_1, \ldots, a_m\}$; $\mathcal{T}(\Gamma)$: l'ensemble des parties de Γ. $\forall a \in \mathcal{T}(\Gamma)\setminus\{\varnothing ; \Gamma\}$, $\forall a_i \in a$, a_i réalise une application de Ω vers {0 ; 1} ou {Faux ; Vrai} et $P(a_i = 1) = Card(a_i^{-1}(1))/n$. Toute partie non vide de Γ sera appelée un motif. Par commodité, dans toute la suite du texte, un motif désignera indifféremment un ensemble d'attributs et un attribut ou une variable. Par souci de congruence phonétique et typographique, nous écrirons, pour deux motifs a et b dans $\mathcal{T}(\Gamma)$: $A = a^{-1}(1)$, $B = b^{-1}(1)$, $n_a = Card(A)$, $n_b = Card(B)$, $n_{ab}=Card(A\cap B)$, $\bar{a}$ désigne la négation de a, et $\bar{A} = \Omega - A$, les réels P(A) et P(B) seront respectivement appelés les support des motifs a et b, notés supp(a) et supp(b).

3. Rappels sur les règles d'association

La notion de règle d'association a été introduite par [1], puis intensivement étudiée par [15], [3], [16]. Nous en rappelons ci-dessous la définition.

Définition 1 : *On appelle règle d'association extraite de la base des données booléennes D, tout couple de motifs* $(a, b) \in \mathcal{T}(\Gamma)^2$, *notée* $a \rightarrow b$ *tel que* $a \cap b = \varnothing$.

Définition 2 : *Une mesure de qualité probabiliste est une fonction réelle* μ *de* $\mathcal{T}(\Gamma)^2$ *telle que pour toute règle d'association* $a \rightarrow b$, $\mu(a \rightarrow b)$ *est calculée à partir des quatre quantités* n, $supp(a)$, $supp(b)$ *et* $supp(a\cap b)$.

Terminologies : Les motifs a et b d'une règle d'association $a \rightarrow b$ seront respectivement appelés son *prémisse* et son *conséquent*. La limitation à ces quatre paramètres, pour définir une mesure de qualité de règles, se justifie par le fait qu'ils suffisent pour retrouver les effectifs correspondants aux cinq cases restantes dans le tableau de contingence obtenu par le croisement de a et b, car $A=(A\cap B)+(A\cap \bar{B})$ et $B=(B\cap A)+(B\cap \bar{A})$. Une règle d'association $a \rightarrow b$ est dite exacte si $A\subset B$ et s'il existe une des mesures de qualité μ_{ai} parmi les mesures considérées telle que $\mu_{ai}(a \rightarrow b) = 1$, sinon elle est dite règle d'association approximative ou une implication partielle [10]. Une règle d'association de type $a \rightarrow \bar{b}$ sera dite une règle négative.

Exemples de mesures de qualités de règles : Les concepts de support et confiance évoqués ci-dessus sont les deux mesures probabilistes de qualité des règles d'association les plus utilisées [12]. Cependant, la considération exclusive de ces deux mesures conduit à deux inconvénients majeurs tels la surabondance des règles générées et l'extraction systématique de certaines règles avérées incohérentes avec la sémantique d'interdépendance statistique tant recherchée dans la pratique. Aussi, bons nombres de travaux alimentent la littérature pour tenter justement de résoudre ces écueils[3] et le problème demeure encore ouvert aujourd'hui. Dans le cadre de l'analyse formelle des concepts, [4] propose un algorithme de génération de la base de Guigues-Duquenne-Luxemburger pour les règles d'associations définies par ces deux mesures de qualité. D'autres approches via des mesures probabilistes de qualité sont également proposées [8], [19]. Remarquons par ailleurs que la plupart des mesures probabilistes de la qualité d'une règle d'association $a \rightarrow b$ peuvent s'exprimer en fonction de la probabilité conditionnelle $P(B \mid A)$. En général, trois facteurs sont pris en compte pour évaluer la qualité d'une règle d'association $a \rightarrow b$: sa couverture égale à cardinal(a,,b), sa confiance, puis sa complétude définie par $P(A \mid \bar{B})$ [5]. C'est dire l'importance du concept de probabilité conditionnelle pour ces types d'indices.

4. Principes de base des mesures de la qualité. Définitions.

Par analogie à l'équivalence logique d'une implication formelle à sa contraposée, nous posons les deux définitions ci-dessous.

Définition 3 : Une mesure μ de qualité de règles d'association est dite mesure d'implication, si pour toute règle d'association $a \rightarrow b$, elle vérifie $\mu(\bar{b} \rightarrow \bar{a}) = \mu(a \rightarrow b)$.

Définition 4 : Une mesure μ de qualité de règles est dite symétrique, si pour toute règle d'association $a \rightarrow b$, on a : $\mu(b \rightarrow a) = \mu(a \rightarrow b)$, et parfaitement symétrique, si pour toute règle d'association $a \rightarrow b$, elle vérifie $\mu(\bar{a} \rightarrow \bar{b}) = \mu(a \rightarrow b)$.

Définition 5 : Une règle d'association dont une des mesures de qualité est implicative sera qualifiée de règle d'association implicative.

Il est facile de vérifier, par exemple, que la mesure support d'une règle d'association est symétrique, mais non implicative, alors que la confiance est non symétrique, mais n'est pas une mesure d'implication ; la mesure de Gras[6] est une mesure d'implication non symétrique. Voici les cinq principes conduisant à notre définition de mesure normée centrée.

(i) Les trois principes de Piatetsky-Shapiro [14]:

Une mesure d'intérêt d'une règle d'association $a \rightarrow b$ doit être nulle en cas d'indépendance statistique des prémisse et conséquent, fonction strictement croissante du nombre n_{ab} d'exemples, les autres paramètres étant fixés, et une fonction strictement décroissante du cardinal n_a du dual de son prémisse a ou décroissant du cardinal n_b du dual de son conséquent, les autres paramètres étant maintenus constants.

(ii) Un quatrième principe de Major & Mangano [11] vient s'ajouter à ceux-là :

Une mesure d'intérêt d'une règle d'association doit être une fonction strictement croissante de sa couverture, une fois que sa confiance est gardée constante supérieure à une valeur minimale préalablement fixée.

(iii) Afin de corriger le caractère symétrique de l'indice de Piatetsky-Shapiro [14], Freitas [5] propose le cinquième principe qu'une mesure de qualité d'intérêt d'une règle d'association doit être non symétrique.

3. Mesure normée centrée de la qualité des règles d'association.

Soit une règle $a \rightarrow b$ et la probabilité conditionnelle $P(B \mid A)$. Nous avons dans [6] les états intuitifs suivants : indépendance statistique entre a et b correspondant à $P(B \mid A)=P(B)$, dépendance positive entre a et b interprétée par une attraction si $P(B \mid A) > P(B)$: on dira alors que a favorise b, dépendance négative interprétée par une répulsion entre a et b si $P(B \mid A) < P(B)$: on dira alors que a défavorise b, incompatibilité correspondant à la probabilité conditionnelle nulle, implication logique si elle atteint la valeur maximale +1 , auquel cas a implique totalement b. En vertu de la continuité de la fonction $P(B \mid A)$ de $1 - \frac{|A \cap B|}{|A|}$, le phénomène d'implication totale est la limite supérieure de la dépendance positive, ou du phénomène d'attraction entre les deux événements en question. Et par dualité, pour rendre compte d'une dépendance négative, dans le cas où l'un des événements défavorise l'autre, une mesure de qualité de règle doit être négative dans ce cas et tendre vers la valeur −1 au cas où les deux événements sont incompatibles. D'où les deux définitions données ci-dessous.

Définition 6 : Une mesure probabiliste $\mu \in R^{\tau \cdot (\tau - 1)}$ de qualité de règle d'association sera dite normée et centrée, si μ vérifie les trois principes de Piatetsky-shapiro, est non symétrique et telle que pour toute règle $a \rightarrow b$ d'un contexte donné, on a : $\mu(a \rightarrow b) > 0$ si a favorise b, $\mu(a \rightarrow b) < 0$ si a défavorise b, $\mu(a \rightarrow b) = +1$ en cas d'implication logique, et $\mu(a \rightarrow b) = -1$ en cas d'incompatibilité. Puisque $n_{a\bar{b}} = n_a - n_{ab}$ pour tout couple (a, b) de motifs ou de variables, tout contre-exemple signifie contradicteur d'implication ici. La proposition suivante est immédiate.

Proposition 1 : Toute mesure normée de la qualité de règles est une fonction strictement décroissante du nombre de contre-exemples sur le conséquent d'une règle présents dans le prémisse de celle-ci.

Terminologie : La mesure probabiliste normée centrée μ_e déduite d'une mesure μ de qualité de règles donnée est appelée la normalisée de μ. Notons que cette définition de mesure normée est différente de celle proposée dans [12], selon laquelle il suffit que le domaine des valeurs de la mesure en question soit l'intervalle [-1, 1] sans prise en compte des sémantiques probabilistes et statistiques évoquées dans la présente. Elle est également différente du concept d'«indice statistiquement normalisé» proposé dans [18] : un indice n'est pas nécessairement une mesure normée centrée.

On vérifie, par exemple, que la mesure Confiance évoquée ci-dessus n'est ni normée, ni centrée car elle vaut zéro au lieu de –1 en cas d'incompatibilité , et ni implicative.

6. La mesure Ion.

Proposition 2 et définition 7

La mesure de qualité, notée Ion, définie par

$Ion(a \to b) = (P(B/A)-P(B))/(1-P(B))$, si a favorise b et $Ion(a \to b) = (P(B/A)-P(B))/P(B)$ sinon [19]

est une mesure d'implication, normée centrée, et n'est pas symétrique. On l'appellera tout simplement Implication statistique orientée normée : on la notera Ion.

En effet : Plaçons-nous dans le cas des deux motifs a et b tels que a favorise b. Dans la suite on symbolisera « a favorise b » par « a f b », et «a défavorise b » par « a df b ».

Alors, $Ion(a \to b) = (n.n_{ab} - n_a.n_b)/n_a(n - n_b)$ est une fonction strictement croissante de n_{ab}, et de $supp(a \to b)$, à marginales et à n fixés. Puis, si A $\subset$B, alors P(B/A)=1, et donc $Ion(a \to b) = (1-P(B))/(1-P(B))=1$. Si B est indépendant par rapport à A, alors P(B/A)=P(B) et donc $Ion(a \to b) = 0$. Dans le cas où a défavorise b, si B est asymptotiquement incompatible avec A, alors $P(B/A) = 0 < P(B)$ implique que $Ion(a \to b) = (0-P(B))/P(B) = -1$.

$Ion(\bar{b} \to \bar{a}) = (P(\bar{A}/\bar{B}) - P(\bar{A}))/(1 - P(\bar{A})) = (P(A)/P(A))(1 - (1 - P(B/A))/(1 - P(B))) = Ion(a \to b)$. Ce qui montre que Ion est une mesure d'implication. Enfin, pour la non symétrie, il est facile de vérifier l'égalité $Ion(a \to b) = Ion(b \to a)$ n'est possible que si l'on se trouve dans un cas particulier où P(B/A) = P(A/B) et P(A)=P(B). D'où Ion est une mesure d'implication, non symétrique, centrée normée. Normalisation et centrage d'une mesure probabiliste : Soit μ une mesure probabiliste de la qualité de règles à normaliser et à centrer, μ_e sa normalisée centrée. Soit $a \to b$ une règle d'association. Désignons par : - x_f et y_f : les coefficients correspondant au cas « a f b », fonction de P(A), P(B) ;

- x_d et y_d : les coefficients correspondant au cas « a df b », fonction de P(A), P(B) .

Compte tenu de la continuité de l'évolution dans les deux zones d'attraction et de répulsion, on a : $\mu_e(a \to b) = x_f.\mu(a \to b) + y_f$, si a f b, et $\mu_e(a \to b) = x_d.\mu(a \to b) + y_d$, si a df b (1).

D'où $x_f = 1/(\mu(a \to b)_{sup} - \mu(a \to b)_{ind})$ et $y_f = \mu(a \to b)_{ind}/(\mu(a \to b)_{sup} - \mu(a \to b)_{ind})$ (2)

$x_d = 1/(\mu(a \to b)_{ind} - \mu(a \to b)_{inc})$ et $y_d = -\mu(a \to b)_{inc}/(\mu(a \to b)_{ind} - \mu(a \to b)_{inc})$ (3)

Ceci montre la possibilité de la normalisation-centrage d'une mesure probabiliste de la qualité des règles. Réciproquement, l'expression de la mesure initiale μ en fonction de sa normalisée centrée μ_e est : $\mu(a \to b) = (\mu_e(a \to b) - y_f)/x_f$ si a favorise b et $(\mu_e(a \to b) - y_d)/x_d$ si a défavorise b (4). Par définition, sa normalisée centrée μ_e évolue selon le schéma ci-dessous (Figure 1).

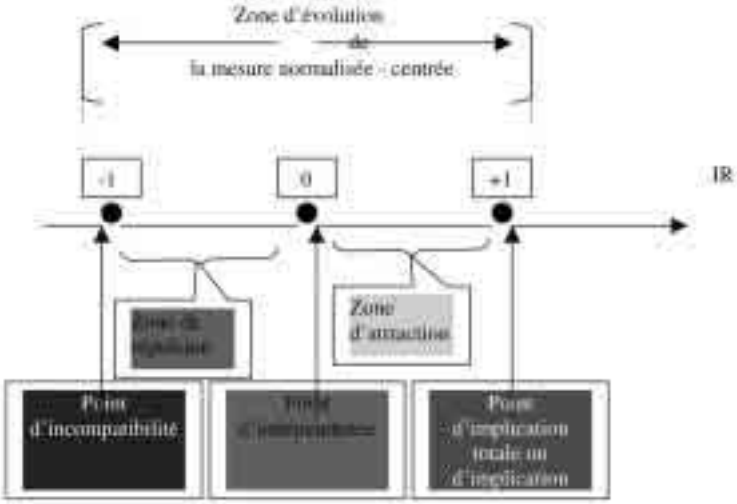

Figure 1 : schéma de normalisation et centrage d'une mesure probabiliste.

Conséquence : Grâce à ces relations réciproques (1) et (4), il sera donc envisageable de comparer les mesures probabilistes normalisables entre elles, leurs normalisées étant toutes égales à Ion (vérification laissée au soin du lecteur). Signalons Ion est une mesure d'implication directement proportionnelle au nombre de contre-exemples sur le conséquent présents dans le prémisse, fonction croissante de $n_{a.\bar{b}}$ comme d'autres mesures implicatives, mais Ion est normée et inversement proportionnelle au nombre total de contre-exemples sur le conséquent. La mesure Ion joue ainsi un rôle central parmi un grand nombre de mesures probabilistes en permettant de les comparer entre elles. Mais en général, l'équivalence n'a pas lieu dans le cas de la mesure normalisée μ_n d'une mesure probabiliste μ.

Néanmoins, on a : $0 \leq \mu_n(a \rightarrow b) < 1 \Leftrightarrow -1 < \mu_n(a \rightarrow \bar{b}) < 0$.

Proposition 3 : Ion est égale à sa propre mesure normalisée centrée et est non symétrique.

En effet, sous l'hypothèse d'indépendance stochastique de a et b, on a : Ion(a → b) = 0, sous l'hypothèse d'implication logique de a sur b, on a Ion(a → b) = (1-P(B))/(1-P(B)) = 1 , sous l'hypothèse d'incompatibilité on a Ion(a → b)=(0-P(B))/P(B) = - 1. La non symétrie peut se démontrer par un simple raisonnement par l'absurde et provient du fait que généralement P(A) ≠ P(B). Voyons sa relation avec certaines mesures probabilistes.

a. Rapport avec la mesure de l'écart à l'indépendance : Par analogie à la distance de Khi-deux, utilisée pour les tests d'indépendance, fonction de (Ob - Th)/Th , où Ob désigne la mesure observée et Th celle théorique calculée sous l'hypothèse d'indépendance stochastique, la quantité définie par Eiba = (P(B/A) – P(B))/P(B) mesure naturellement l'écart à l'indépendance de la variable b par rapport à a, si a favorise b. Or Ion(a → b) = P(B)/(1 – P(B)) > Eiba. D'où la proposition 4. Néanmoins, on a Ion(a → b) = Ei $\overline{\text{ba}}$.

Proposition 4 : Si les deux variables a et b sont telles que P(B)>0,5 et P(B/A)>P(B), alors Ion(a → b) est supérieur à l'écart à l'indépendance. Ainsi la mesure normée Ion est un indicateur de réduction de l'incertitude de b sachant la réalisation de a.

La relation explicite entre Ion et Khi-deux donnée ci-après renseigne sur l'allure de la distribution statistique de Ion

b. Concernant les règles négatives, on a :

Proposition 5 : Pour toute règle $a \to b$, on a : $Ion(a \to \overline{b}) = -Ion(a \to b)$. (5)

Proposition 6 : $\forall (a \to b), \forall \alpha \in]0,1[, \alpha \leq Ion(a \to b) < 1 \Leftrightarrow -1 < Ion(a \to \overline{b}) \leq -\alpha$

c. Ion et coefficient de corrélation et Khi-deux d'indépendance, sa distribution :

Proposition 7 : Pour toute règle $a \to b$, on a : $Ion(a \to b) > 0 \Leftrightarrow$ corrélation $(a, b) > 0$.

En effet : $Ion(a \to b) = \sqrt{(n_b n - n_a)/(n_a n - n_a n_b)}\,\phi(a \to b)$. Donc Ion et ϕ ont constamment le même signe. De plus, de $\chi^2 = n\,\phi^2(a \to b)^2$ résulte sa relation directe avec la statistique de χ^2 : $Ion(a \to b) = \pm\sqrt{(1/n).(n_a/n_b).(n - n_a)/(n - n_b)}.\chi^2$ (6)

Par conséquent, comme dans [17], à partir de la table de Khi-deux, de (6) résulte une abaque des valeurs critiques de Ion pour décider sur sa signification selon les supports respectifs de a et de b. Par exemple, dans le cas où Confiance$(a \to b) >$ Confiance$(\overline{b} \to a)$ (dépendance positive de a sur b), on a la correspondance suivante : au niveau de signification de 95%, la valeur critique de χ^2 étant 3,84, celle de ϕ vaut $1,96/\sqrt{n}$, celle de $Ion(a \to b)$: $\sqrt{(n_b/n_a).(n - n_a/n - n_b)}.(1,96/n)$. Par rapport à la mesure de la surprise [2] définie par : $Surp(a \to b) = (P(A \cap B) - P(A \cap \overline{B}))/P(B)$ (7). On montre le résultat suivant.

Proposition 8 : $Surp_+ = Ion_+$ et Si $Ion(a \to b) > 0$ et $P(B) \leq 0,5$, alors $surp(a \to b) > 0$.

Corollaire : La mesure Ion est une mesure d'implication, non symétrique, centrée, normée, orientée positivement ou négativement selon la dépendance statistique, et qui prend en compte le degré surprise.

6. Algorithme d'extraction proposé :

Si $Ion(a \to b)$ est significatif et $confiance(a \to b) > 1/2$, alors retenir la règle $(a \to b)$ avec sa mesure normalisée et sa confiance ; Sinon Si $Ion(a \to \overline{b})$ est significatif et $confiance(a \to b) > 1/2$, alors retenir la règle négative $(a \to \overline{b})$ et sa confiance. Sinon pas de règle générée, à un certain seuil de confiance fixé par l'utilisateur. Fin.

7. Traitement de variables latticielles

Dans ce paragraphe nous mon montrons comment il est possible d'étendre les résultats précédents à des variables non nécessairement binaires [19].

Soit $\Gamma = (a_i)_{i \in I}$ un ensemble de variables binaires, notons $B = [V, F]$. La variable binaire a_i est une application $a_i \in B^{\Omega}$. Considérons B comme une chaîne $B = [F < V]$, l'ensemble B^{Ω} est alors ordonné tel que : $a_i \leq a_k \Leftrightarrow \forall \omega \in \Omega,\ a_i(\omega) \leq a_k(\omega)$. On définit l'inf $a_i \wedge a_k$ de deux variables comme la variable binaire : $\forall \omega \in \Omega,\ a_i \wedge a_k(\omega) = a_i(\omega) \wedge a_k(\omega)$ (8)

Soit $M = T(\Gamma)$ l'inf-demi-treillis engendré par Γ. Un motif est représenté par un élément $a = \wedge_{k \in K} a_k \in M$, avec $K \subset I$, appelons extension d'un motif, l'ensemble des observations $\omega \in \Omega$ reconnu par le motif : $A = a^{-1}(V) = ext(a)$ (i.e. extension de a).

Le support d'un motif a est : $supp(a) = P(A) = card(A)/card(\Omega) = card(ext(a))/n$ (9)

Une règle d'association est un couple de motifs $a = \wedge_{k \in K} a_k \in M$ et $b = \wedge_{h \in L} b_h \in M$, avec $K \subset I$ et $L \subset I$, noté $(a = \wedge_{k \in K} a_k \to b = \wedge_{h \in L} b_h) \in M^2$ ou $(a \to b)$, vérifiant $I \cap K = \emptyset$ et $supp(a).supp(b) \neq 0$.

Une mesure de qualité d'une règle d'association μ est une fonction réelle définie sur M^2. Elle est dite probabiliste si μ ne dépend que de $(supp(a), supp(b), supp(a \wedge b), n = card(\Omega))$.

Notons les motifs complémentaires $\overline{a} = \wedge_{k \in \overline{K}} a_k$, avec $\overline{K} = J - K$ et de même $\overline{b} = \wedge_{h \in \overline{L}} b_h$, avec $\overline{L} = J - L$. Le tableau de contingence pour le calcul de la mesure de qualité $Ion(a \to b)$ est $C = (c_{uv})$ où $c_{uv} = card(ext(a_u \wedge v))$, avec $u \in \{a, \overline{a}\}$ et $v \in \{b, \overline{b}\}$. Remarquons que l'on peut identifier chaque observation ω de Ω avec un élément (appelé

item) de M. Il suffit de considérer le plus grand ensemble $K_\omega \subset J$ tel que $\forall a \in K_\omega : a_k(\omega) = V$, alors à ω on associe $\wedge_{k \in K_\omega} a_k \in M$. Dès lors on peut considérer des variables qui ne sont pas simplement binaires

pour la recherche de règles d'association caractérisées par la mesure de qualité lon. Soit l'ensemble des observations Ω défini sur un ensemble de variables Γ induisant une structure de inf-demi-treillis : $\Omega \subset M = T(\Gamma)$. Tout $a \in M$ peut être considéré comme une fonction binaire $a: B\Omega$ de la manière suivante : $a(\omega) = V$ si $\omega \leq a$ sinon $a(\omega) = F$. Les concepts d'extension, de règles d'association et de support s'en déduisent facilement ainsi que la mesure de qualité lon. Par exemple, Γ peut contenir :

– des variables qualitatives : $Q = \{a_j\}_{j \in K}$ où les modalités $a_j \in B^\Omega$ vérifient $a_j \wedge a_k = \perp$ si $k \neq j$, la fonction $\perp$ étant telle que $\perp(\omega) = F$.

– des variables hiérarchiques : $H = \{a_i\}_{i \in K}$ où chaque a_i est une feuille ou bien est de la forme $a_i = \vee a_l$

– des variables intervalles (voir [19] pour le détail).

8. Discussion et perspectives

Sachant qu'une probabilité discrète P(A) d'un événement se généralise par une densité de probabilité dans le cas d'une variable aléatoire absolument continue, ne serait-il pas naturel d'adopter la définition suivante : Si X et Y sont deux variables aléatoires à densités respectives f_X et f_Y, telles que le vecteur aléatoire (X, Y) a pour densité $f_{(X, Y)}$, on dira que la variable aléatoire X implique Y, si pour tous réels x et y, on a l'inégalité $f(y/x) \geq f_Y(y)$?

Cependant, se pose la question de la sémantique d'un tel concept d'implication entre variables continues : quelle interprétation peut-on y donner ? Nous traitons cette problématique dans un prochain rapport. Néanmoins, d'ores et déjà, il est loisible d'utiliser la mesure normée centrée lon pour définir et décider d'éventuels liens d'implication entre certaines classes de variables dans le cadre d'une analyse classificatoire : pour deux classes C1 et C2 formées, à l'instar de l'implication entre variables booléennes, il suffit de raisonner à partir du tableau de contingence obtenu par leur croisement à partir de la base des données étudiées et d'appliquer ensuite la mesure lon. Par ailleurs, il semblerait que l'ensemble des mesures de similarité comporte une classe des mesures d'implication ou des mesures de similarité implicative.

9. Conclusion

Cette étude montre d'une part, l'existence d'une mesure normée et centrée lon, non symétrique, qui joue en fait un rôle central par rapport aux indices probabilistes usuels pour apprécier la qualité des règles d'association avec dépendance orientée interprétable en terme d'implication statistique, cette mesure ayant un rôle comparable à celui de la loi normale centrée et réduite au sein des lois gaussiennes dans le champs des variables aléatoires : lon permet d'identifier si une règle est valide ou non, simplement par la connaissance de la valeur de sa mesure, même si celle-ci ne prend pas les valeurs constantes $-\mathcal{L}$, 0, $+\mathcal{L}$ respectivement aux états d'incompatibilité, d'indépendance, d'implication logique. D'autre part, le présent travail, encore relativement théorique certes, offre aussi l'opportunité d'une approche probabiliste multicritère pour extraire des règles d'associations plus cohérentes avec la notion de dépendance et de surprise statistiques. C'est une autre preuve de l'insuffisance des deux seuls indices de support et confiance dans la fouille des règles d'association dans une stratégie d'évitier la surabondance de telles règles extraites. Le travail de validation par l'application à un jeu des données réelles à la suite de l'élaboration d'un logiciel opérationnel est en cours dans notre laboratoire. Par ailleurs, l'inexistence d'un critère meilleur que tous les autres dans cette tâche de fouille des règles étant acquise, il semblerait à travers cette étude qu'il existerait des jeux de combinaison de critères ou contraintes selon certaines classes de contextes de l'utilisateur expert de ses données ; de plus un invariant semblerait incontournable dans ces combinaisons efficaces qui restent encore à identifier.

10. Bibliographie

[1] R. Agrawal, T. Imielinski & A. Swami , 1993. Mining association rules betwen sets of items in large databases. In P. Buneman und S. Jajodia, editors, Proc. of *ACM SIGMOD International Conference on Management of Data*, volume 22, pp. 207-216, Wachington, 1993. ACM press.

[2] Azé Jérome, 2003. Une nouvelle mesure de qualité pour l'extraction de pépites de connaissances, Extraction des connaissances et apprentissage *RSTI série RIA-ECA, Vol. 17-* n°1-2-3-/2003, Extraction et gestion des connaissances EGC 2003, 171-182.

[3] Brin S., Motwani R. & Ullman J.D., & Tsur S, 1997. Dynamic itemset counting and implications rules for market basket data. Proc. Of the *1997 ACM SIGMOD conf, mai 1997b*, 255-264.

[4] J. Diatta , 2003. Génération de la base de Guigues-Duquenne-Luxemberger pour les règles d'association par une approche utilisant des mesures de similarité multivoies. In *Conf. d'apprentissage*, Laval, France, 2003, p. 281-298. Presse universitaire de Grenoble.

[5] Freitas A.A., 1999. On rule of interestingness measure. *Knowledge –Based-Systems n°12*, 1999, 309-315.

[6] Gras R., Totohasina A., 1995. Chronologie et causalité, sources d'obstacles épistémologiques à l'apprentissage de la notion de probabilité conditionnelle. *Recherche en didactique des mathématiques, vol.15*, n°1, p.49-95, 1995 Edit° La Pensée sauvage, France.

[7] Lesca Philipe, Meyer Patrick, philippe Picouet, Vaillant Benoît, Lallich Stéphane, 2003. Critères d'évaluation des mesures de qualité en ECD. *Société française de statistique, Actes des XXXVèmes Journées de Statistique, T.2*, Lyon, 2-6 juin 2003 , 647-650.

[8] Lerman I.C., Azé J., 2003, Une mesure probabiliste contextuelle discriminante de qualité des règles d'association., *RSTI s. RIA-ECA, Vol. 17-* n°1-2-3-/03, EGC, 247-262.

[9] Luxemburger M., 1991. Implications partielles dans un contexte. *Mathématiques Informatique Sciences humaines, 29ième année, n°113*, 1991, 35-55.

[10] J.A. Major and J.J. Mangano, 1993. Selecting among rules induced from a hurricane database. In *KDD'93*, Workshop papers, pages 28–41, Menlo Park, California.

[11] Hilderman Robert J et.Hamilton Howard J. (1999). Knowledge discovery : a survey. Tehnical report, Dept of Computer Sci., university of Regina, Canada S4S 0A2, 1999

[12] J. Hipp, U. Güntzer & G. Nakaeizadeh, 2000. Algorithm for Association Rule Mining – A general Survey and Comparison. *SIGKDD Explorations, Vol. 2*, 58-64.

[13] Piatetsky-shapiro G., 1991. Knowledge discovery in Real Data Bases : A report on the *IJCAI-89* Workshop, AI Magazine, 11(5), 1991, 68-70.

[14] Tan Pang-Ning, Kumar Vipin & Srivastava Jaideep, 2002. Selecting the right interestingness measure for association patterns. Proc. *SIGKDD'02*, Edmonton, Alberta.

[15] A. Savasere, E. Omiecinski, S. Naathe, 1995. An efficient algorithm for mining association rules in large databases. In Proc. Of *the 21th VLDB Conference*, 432-444.

[16] Totohasina A., 2003. Normalisation des mesures probabilistes de la qualité des règles d'association. *Société française de statistique, Actes des XXXVièmes Journées de statistiques, Tome 2*, Lyon, 2-6 juin 2003, pp.985-988.

[17] Totohasina A., 1993 : Notes sur l'implication statistique en classification. Rapport technique du SCAD, Département de maths et informatique, Université du Québec à Montréal, Canada.

[18] Totohasina A., Ralambondrainy H., Diatta J., 2003 : Un algorithme efficace d'extraction des règles d'association implicative. Equipe ECD–IREMIA, Dépt de maths-Info, Université de La Réunion, France.

[19] M.J. Zaki & M.Ohihara, 1998. Theoritical foundations of association rules. In *3rd SIGMOD'98* workshop on Research Issues In data Mining and Knowledge Discovery (DMKD), 1-8, 1998.