Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel TIC

Julien Blanchard, Fabrice Guillet, Régis Gras, Henri Briand

IRIN – Ecole polytechnique de l'université de Nantes
La Chantrerie – BP 50609 – 44306 Nantes cedex 3
{julien.blanchard, fabrice.guillet, regis.gras, henri.briand}@polytech.univ-nantes.fr

Résumé. La validation des connaissances est l'une des étapes les plus problématiques d'un processus de découverte de règles d'association. Pour que le décideur (expert des données) puisse trouver des connaissances intéressantes dans les grandes quantités de règles produites par les algorithmes de fouille de données, il est nécessaire de mesurer la qualité des règles. Nous insérant dans le cadre de l'analyse statistique implicative, nous proposons dans cet article d'évaluer les règles en considérant leur contenu informationnel à travers un nouvel indice de qualité fondé sur l'entropie de Shannon : *TIC* (*Taux Informationnel modulé par la Contraposée*). Cet indice a l'avantage d'être bien adapté à la sémantique des règles, puisque d'une part il respecte leur caractère asymétrique et d'autre part il tire profit de leurs contraposées. Par ailleurs, c'est à notre connaissance la seule mesure de qualité de règles qui intègre à la fois indépendance et déséquilibre, c'est-à-dire qui permette de rejeter simultanément les règles entre variables corrélées négativement et les règles qui possèdent plus de contre-exemples que d'exemples. Des comparaisons de *TIC* avec la J-mesure, l'information mutuelle, l'indice de Gini, et la confiance sont réalisées sur des simulations numériques.

1. Introduction

Parmi les modèles de connaissances utilisés en Extraction de Connaissances dans les Données (ECD), les règles d'association [Agrawal *et al.*, 1993] sont devenues un concept majeur qui a donné lieu à de nombreux travaux de recherche. Ces règles sont des tendances implicatives de la forme $a \rightarrow b$ entre les attributs d'une base de données (variables booléennes dénommées items). Une telle règle signifie que la plupart des enregistrements qui vérifient la prémisse a dans la base de données vérifient aussi la conclusion b. L'une des étapes les plus problématiques d'un processus de découverte de règles d'association est la validation des règles après leur extraction. Les algorithmes de *data mining* peuvent en effet produire d'énormes quantités de règles, ce qui empêche le décideur (expert des données étudiées) de pouvoir exploiter les résultats directement à la sortie des algorithmes. Ce problème est lié à la nature non supervisée de la découverte de règles : le décideur n'explicite pas ses buts et ne spécifie aucune variable endogène. Le nombre de conjonctions d'items manipulées par les algorithmes devient alors prohibitif.

Pour assister le décideur dans sa recherche des connaissances intéressantes pour la prise de décision, il est nécessaire de mesurer la qualité des règles extraites. Dans l'importante littérature consacrée à l'évaluation de cette notion complexe de qualité, les mesures sont souvent classées en deux catégories : les subjectives (orientées décideur) et les objectives

(orientées données). Les mesures subjectives prennent en compte les objectifs du décideur et ses connaissances *a priori* sur le domaine étudié (voir [Padmanabhan et Tuzhilin, 1998], [Liu *et al.*, 1999] pour une vue d'ensemble), tandis que seules les cardinalités liées à la contingence des données interviennent dans le calcul des mesures objectives (voir par exemple [Bayardo et Agrawal, 1999], [Tan *et al.*, 2002], [Hilderman et Hamilton, 2001]). Ces dernières mesures sont nombreuses et de natures diverses : on trouve des mesures fréquentielles, des mesures dérivées de la théorie de l'information, ou bien des mesures fondées sur des tests statistiques. Selon qu'elles sont symétriques (invariantes par permutation de la prémisse et de la conclusion) ou non, elles évaluent des similarités ou des règles.

Dans cet article, nous nous intéressons aux mesures objectives. Ce travail s'insère dans le cadre de l'analyse statistique implicative proposé par Gras [Gras, 1996] et dans le prolongement de l'intensité d'implication entropique [Gras *et al.*, 2001]. Selon nous, la qualité objective d'une règle réside en trois notions :

- sa *généralité*, qui peut être mesurée par le support [Agrawal *et al.*, 1993], le support causal [Kodratoff, 2001] ;
- sa *puissance implicative* (c'est-à-dire la validité de l'inclusion sous-jacente $A \subseteq B$ pour une règle $a \rightarrow b$ [1]), qui peut être mesurée par la confiance [Agrawal *et al.*, 1993], l'indice de Loevinger [Loevinger, 1947], ou la J-measure [Smyth et Goodman, 1991] pour ne citer qu'elles parmi les mesures non symétriques ;
- sa *significativité statistique*, qui peut être mesurée par un test du χ^2 [Brin *et al.*, 1997a], l'intensité d'implication [Gras, 1996], ou sa version entropique [Gras *et al.*, 2001 ; Blanchard *et al.*, 2003a] qui intègre à la fois significativité et puissance implicative.

Avec ces trois critères, le décideur peut privilégier différents sous-ensembles de règles. Par exemple, si le décideur désire exploiter des "pépites" dans les données, telles des niches comportementales, il va rechercher des règles *spécifiques* mais *significatives*, et en contrepartie tolérera une *puissance implicative* plus faible (plus de contre-exemples). Par contre si le décideur s'intéresse à des connaissances bien connues, par exemple pour conforter une théorie ou mettre en confiance un expert des données, ou bien à des fins pédagogiques, alors il recherchera plutôt des règles *générales*, *significatives* et de *puissances implicatives* élevées. A l'extrême, si le décideur ne recherche pas d'éventuels effets de causalité mais désire une description des données, il s'appliquera à trouver des règles fortement *implicatives* mais qui ne sont pas forcément *significatives*.

Nous proposons dans cet article d'évaluer la puissance implicative des règles à l'aide d'un indice fondé sur l'entropie de Shannon [Shannon et Weaver, 1949] : le *taux informationnel*. Il a l'avantage de posséder une sémantique claire, puisqu'il mesure la quantité d'information (gain d'entropie) apportée par la règle. Cette propriété sémantique est primordiale pour un bon indice de qualité afin que le décideur puisse sélectionner des mesures en lesquelles il a confiance. Il existe d'autres mesures issues de la théorie de l'information communément utilisées pour évaluer la qualité des règles : le taux d'information mutuelle [Tan *et al.*, 2002 ; Jaroszewicz et Simovici, 2001], la J-mesure [Smyth et Goodman, 1991], l'indice de Gini [Bayardo et Agrawal, 1999], et la mesure de surprise de Freitas [Freitas, 1999]. Pour une règle $a \rightarrow b$, l'information mutuelle (gain d'entropie de Shannon) mesure l'information

[1] Voir les notations définies partie 2.

moyenne partagée entre les variables *a* et *b*, et la J-mesure est la part de l'information mutuelle relative à la règle, ou plus précisément aux évènements (*a* et *b*) et (*a* et $\bar{b}$). L'indice de Gini est quant à lui issu de l'entropie quadratique. Enfin, la mesure de Freitas est un indicateur de l'effet de surprise provoqué par une règle. Elle repose sur un paradigme différent des mesures précédentes puisqu'elle cherche à minimiser l'information apportée sur la conclusion par chaque attribut en prémisse, et non à la maximiser. En effet, une règle bien évaluée par d'autres mesures est d'autant plus étonnante que les attributs qui la constituent sont, pris indépendamment les uns des autres, peu informatifs pour la conclusion.

La plupart des mesures de qualité des règles ne prennent pas en compte les règles contraposées[2]. Pourtant, l'utilité de la contraposée pour la découverte de règles a été explicitée dans [Kodratoff, 2001] : si la sémantique de la relation recherchée dans les données est descriptive, alors la contraposée ne doit pas être considérée (comme le suggère le paradoxe de Hempel) ; par contre si la sémantique de la relation recherchée est causale, alors la contraposée doit être prise en compte (le paradoxe n'a pas cours dans ce cas). Considérant que le décideur recherche intuitivement des règles de nature causale (le modèle de référence est l'implication logique), nous avons choisi dans cet article de faire valoir la contraposée pour confirmer ou infirmer la puissance implicative des règles (approche que nous avions déjà adoptée avec l'intensité d'implication entropique). Pour cela, nous introduisons dans la partie 2 la notion de *couple implicatif*.

Dans cet article, nous présentons une nouvelle mesure objective de qualité qui évalue la puissance implicative des règles. Cette mesure, appelée *TIC*, est fondée sur le gain d'entropie. Elle permet de repérer à la fois l'*indépendance* (corrélation nulle entre les variables prémisse et conclusion) et le *déséquilibre* (autant d'exemples que de contre-exemples) des règles. Nous introduisons dans la partie suivante la notion de taux informationnel *TI* d'une règle puis définissons la mesure *TIC* qui prend en compte conjointement la règle et sa contraposée. Dans la partie 3, nous étudions les propriétés de ces deux indices et réalisons des comparaisons de *TIC* avec la J-mesure, l'information mutuelle, l'indice de Gini, et la confiance sur des simulations numériques.

2. Taux informationnel d'une règle

Nous considérons un ensemble *T* de *n* sujets décrits par un ensemble *I* de variables booléennes. Dans le vocabulaire des règles d'association, les sujets sont des transactions stockées dans une base de données, les variables sont appelées des items et les conjonctions de variables des itemsets. Etant donnée un itemset *x*, nous notons *X* l'ensemble des transactions qui vérifient *x* et $\overline{X}$ le complémentaire de *X* dans *T*. La cardinalité d'un ensemble *X* est notée n_X. Une règle d'association est un couple (*a*, *b*) noté $a \rightarrow b$ où *a* et *b* sont deux itemsets qui ne possèdent pas d'item en commun. Nous définissons un *couple implicatif* noté $a \Rightarrow b$ comme étant la paire de règles $\{a \rightarrow b,\ \bar{b} \rightarrow \bar{a}\}$. Dans la suite, nous appelons "variables" les itemsets.

[2] Le plus souvent, les mesures qui évaluent "plus" qu'une simple règle prennent en compte la réciproque $b \rightarrow a$ de la règle $a \rightarrow b$, considérant ainsi les règles comme des similarités.

2.1. Une fonction entropique non symétrique

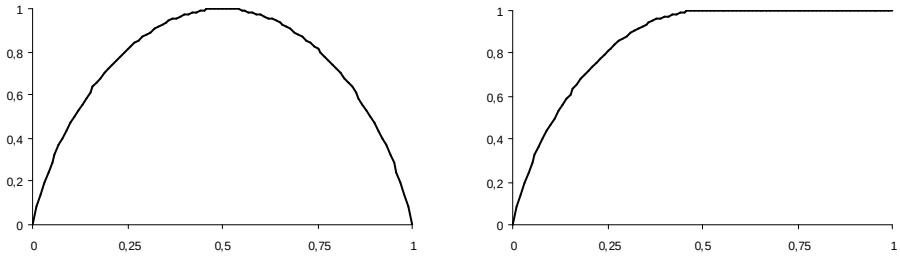
Une approche couramment adoptée dans les indices de qualité pour évaluer la puissance implicative des règles $a \to b$ est de quantifier le déséquilibre entre les exemples $A \cap B$ et les contre-exemples $A \cap \overline{B}$. Pour mesurer ce déséquilibre, nous considérons l'expérience aléatoire qui consiste à vérifier si *b* est réalisé quand *a* est observé. L'incertitude moyenne de l'expérience est mesurée par l'entropie conditionnelle $E_{b/a=1}$ de la variable *b* sachant la réalisation de *a* :

$$E_{b/a=1} = -\frac{n_{A\cap B}}{n_A}\log_2\frac{n_{A\cap B}}{n_A} - \frac{n_{A\cap \overline{B}}}{n_A}\log_2\frac{n_{A\cap \overline{B}}}{n_A}$$

C'est sur cette entropie qu'est fondé l'indice d'inclusion, mesure de puissance implicative des règles qui intervient dans le calcul de l'intensité d'implication entropique [Gras *et al.*, 2001 ; Blanchard *et al.*, 2003a]. Cependant, à l'instar de l'information mutuelle en théorie de l'information, il peut s'avérer judicieux de mesurer le gain d'entropie apporté par la règle, c'est-à-dire l'écart entre l'entropie *a posteriori* précédente et l'entropie *a priori* de la variable *b*. Ce gain d'entropie est la quantité moyenne d'information apportée par la réalisation de la prémisse au sujet de la conclusion, soit en d'autres termes la quantité d'information contenue dans la règle. Pour une règle $a \to b$, le gain d'entropie ou gain informationnel s'écrit :

$$G(a \to b) = E_b - E_{b/a=1} \qquad \text{où} \qquad E_b = -\frac{n_B}{n}\log_2\frac{n_B}{n} - \frac{n_{\overline{B}}}{n}\log_2\frac{n_{\overline{B}}}{n}$$

Par sa symétrie, l'entropie E_b est autant relative à *b* qu'à $\overline{b}$. De la même façon, l'entropie $E_{b/a=1}$ évalue de la même façon un déséquilibre en faveur de $n_{A\cap B}$ ou un déséquilibre en faveur de $n_{A\cap \overline{B}}$. Cependant seul le déséquilibre en faveur des exemples $n_{A\cap B}$ respecte le sens de la règle $a \to b$ et apporte l'information escomptée, tandis que le déséquilibre en faveur des contre-exemples $n_{A\cap \overline{B}}$ traduit la règle $a \to \overline{b}$. Considérant que la règle n'a pas de signification implicative lorsque les contre-exemples sont plus nombreux que les exemples, nous introduisons une version modifiée de l'entropie qui vaut 1 (incertitude maximale) lorsque le déséquilibre n'est pas orienté convenablement. Cette fonction entropique non symétrique est notée $\hat{E}(x)$ et représentée figure 1.



entropie de Shannon pour une variable binaire
$E(x) = -x\log_2(x) - (1-x)\log_2(1-x)$

entropie réduite $\hat{E}(x)$

FIG. 1 – *Représentations de l'entropie de Shannon et de l'entropie réduite*

Définition 1. L'*entropie réduite* $\hat{E}(x)$ est définie par :

- si $x = 0$, alors $\hat{E}(x) = 0$
- si $x \in]0 ; 0{,}5]$, alors $\hat{E}(x) = -x\log_2(x) - (1-x)\log_2(1-x)$
- si $x \in]0{,}5 ; 1]$, alors $\hat{E}(x) = 1$

2.2. Définition d'un gain informationnel normalisé ou taux informationnel *TI*

Afin de respecter le caractère asymétrique des règles, nous utilisons l'entropie réduite $\hat{E}(x)$ pour mesurer les gains informationnels. Le gain informationnel $\hat{G}$ de la règle $a \rightarrow b$ s'écrit maintenant :

$$\hat{G}(a \rightarrow b) = \hat{E}_b - \hat{E}_{b/a=1} \qquad \text{avec} \quad \hat{E}_b = \hat{E}\left(\frac{n_{\overline{B}}}{n}\right) \quad \text{et} \quad \hat{E}_{b/a=1} = \hat{E}\left(\frac{n_{A\cap\overline{B}}}{n_A}\right)$$

Plus ce gain est important, plus la réalisation de *a* apporte d'information sur *b* et plus la puissance implicative de la règle est garantie. Si le gain est négatif, cela signifie que la règle n'apporte aucune information sur la connaissance de *b*, et même qu'elle en "retire". En d'autres termes, l'incertitude est moindre à prédire *b* sans connaissance *a priori* (au hasard) qu'à prédire *b* en utilisant la règle.

Le gain informationnel est maximal lorsque la règle n'admet aucun contre-exemple. Pour faciliter le filtrage des règles les plus informatives, nous normalisons le gain informationnel en associant le score maximal de 1 aux règles les meilleures. Cela revient à calculer le taux de réduction de l'entropie.

Définition 2. Nous définissons le *Taux Informationnel TI* d'une règle $a \rightarrow b$ par :

$$TI(a \rightarrow b) = 1 - \hat{E}\left(\frac{n_{A\cap\overline{B}}}{n_A}\right) / \hat{E}\left(\frac{n_{\overline{B}}}{n}\right) \qquad \text{si} \quad n_{\overline{B}} \neq 0$$

La mesure n'est pas définie si $n_B = n$, mais ces règles sont évidemment à rejeter (le gain informationnel est d'ailleurs nul).

2.3. Prise en compte de la contraposée au sein de la mesure *TIC*

Afin que les règles contraposées jouent leur rôle dans l'évaluation de la puissance implicative des règles, nous associons le taux informationnel de la règle avec celui de sa contraposée au sein d'un indice synthétique. La solution la plus naturelle pour agréger des quantités d'information consiste à utiliser la moyenne arithmétique, comme dans l'information mutuelle moyenne entre deux variables. Cependant, il nous semble important que l'indice soit nul dès que la puissance implicative de la règle ou de sa contraposée n'existe pas. Nous avons donc choisi d'utiliser la moyenne géométrique pour combiner les taux informationnels de la règle et de sa contraposée en rejetant tous les taux informationnels négatifs.

Définition 3. Le *Taux Informationnel modulé par la Contraposée TIC* d'un couple implicatif $a \Rightarrow b$ est la moyenne géométrique des taux informationnels des règles $a \rightarrow b$ et $\bar{b} \rightarrow \bar{a}$. Il est défini par :

- $TIC(a \Rightarrow b) = \sqrt{TI(a \rightarrow b) \times TI(\bar{b} \rightarrow \bar{a})}$ si $TI(a \rightarrow b) \geq 0$ et $TI(\bar{b} \rightarrow \bar{a}) \geq 0$
- $TIC(a \Rightarrow b) = 0$ sinon

TIC évalue non pas une simple règle $a \rightarrow b$ mais un couple implicatif $a \Rightarrow b$. Cependant, dans la suite, afin d'employer un vocabulaire commun avec celui des autres mesures de qualité, nous évoquons indifféremment le *TIC* d'un couple implicatif ou le *TIC* d'une règle.

Les valeurs élevées de l'indice *TIC* mettent en évidence des couples implicatifs dont la règle directe et/ou sa contraposée ont une forte puissance implicative. Toutefois, dans le cadre d'une évaluation des règles fortement sélective, il peut être judicieux plutôt que de calculer une moyenne de retenir le minimum des deux taux informationnels. Les valeurs élevées de *TIC* sont alors attribuées à des couples implicatifs très forts, dont les règles sous-jacentes possèdent toutes les deux une forte puissance implicative. Cela permet en particulier d'écarter les règles qui, aux yeux du décideur, pourraient être confirmées trop fortement par la contraposée et trop faiblement par la règle directe.

3. Propriétés des mesures *TI* et *TIC*

3.1. Taux informationnel *TI*

TI prend ses valeurs dans $]1-1/\hat{E}(n_{\bar{B}}/n)$; 1]. Il n'associe pas les mêmes valeurs à une règle $a \rightarrow b$ et à sa contraposée, à sa réciproque $b \rightarrow a$, ou à sa règle contraire $a \rightarrow \bar{b}$. *TI* est une fonction décroissante convexe du nombre de contre-exemples. Il fait partie des indices de qualité "exigeants" qui diminuent rapidement dès les premiers contre-exemples. Plus précisément :

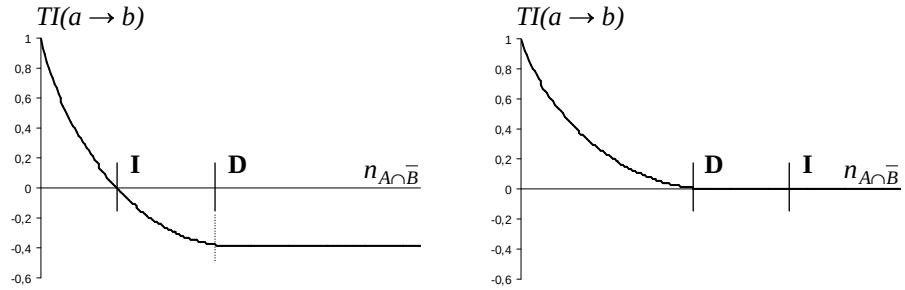
- $TI(a \rightarrow b) = 1$ si $n_{A \cap \bar{B}} = 0$ (implication logique) ;
- $TI(a \rightarrow b) = 0$ si $n_{A \cap \bar{B}} = n_A \times n_{\bar{B}} / n$, ce qui correspond aux cardinalités attendues sous hypothèse d'indépendance entre *a* et *b* ;
- $TI(a \rightarrow b) = 1-1/\hat{E}(n_{\bar{B}}/n)$ au-delà du déséquilibre exemple/contre-exemples, c'est-à-dire lorsque $n_{A \cap \bar{B}} \geq n_A/2$.

On peut distinguer deux comportements différents pour le taux informationnel selon que l'*indépendance* ($n_{A \cap \bar{B}} = n_A \times n_{\bar{B}} / n$) est atteinte avant ou après le *déséquilibre* ($n_{A \cap \bar{B}} = n_A/2$) quand les contre-exemples augmentent :

- si $n_B \geq n/2$, alors l'indépendance est atteinte avant le déséquilibre ; le taux s'annule puis admet des valeurs négatives (figure 2.a).
- si $n_B \leq n/2$, alors le déséquilibre est atteint avant l'indépendance ; le taux s'annule mais n'admet pas de valeurs négatives (figure 2.b). Il est en effet impossible que la règle "retire" de l'information puisque l'incertitude sur *b* est déjà maximale ($\hat{E}_b$=1).

TI permet de repérer à la fois les situations d'indépendance, où les règles ne sont pas significatives, et les situations de déséquilibre, où les règles n'ont pas le sens implicatif escompté. En effet, en ne retenant que les taux informationnels strictement positifs (règles informatives), le décideur rejette l'indépendance, et ainsi écarte les règles entre variables

corrélées négativement, mais rejette aussi le déséquilibre, et ainsi écarte les règles qui possèdent plus de contre-exemples que d'exemples. Parmi les mesures de qualité pour les règles, certaines prennent une valeur fixe (indépendante des données) au déséquilibre mais ne permettent pas de rejeter l'indépendance à l'aide d'un seuil fixe (la confiance [Agrawal *et al.*, 1993], la mesure de Sebag et Schoenauer [Sebag et Schoenauer, 1988], la surprise [Azé et Kodratoff, 2001], l'indice d'inclusion [Gras *et al.*, 2001 ; Blanchard *et al.*, 2003a]…), tandis que d'autres prennent une valeur fixe à l'indépendance mais ne permettent pas de rejeter le déséquilibre à l'aide d'un seuil fixe (le lift [Brin *et al.*, 1997a], l'indice de Loevinger [Loevinger, 1947], la conviction [Brin *et al.*, 1997b], l'indice de Piatetsky-Shapiro [Piatestsky-Shapiro, 1991]…). A notre connaissance, *TI* est le seul indice qui intègre à la fois indépendance et déséquilibre.



(a) cas où $n_B \geq n/2$ (b) cas où $n_B \leq n/2$

FIG. 2 – *Evolution du taux informationnel TI en fonction des contre-exemples*
(I désigne l'indépendance $n_{A\cap\overline{B}} = n_A \times n_{\overline{B}} / n$ *et D le déséquilibre* $n_{A\cap\overline{B}} = n_A/2$*)*

3.2. Taux informationnel modulé par la contraposée *TIC*

TIC prend ses valeurs dans [0 ; 1]. Il n'associe pas les mêmes valeurs à une règle $a \rightarrow b$ et à sa réciproque $b \rightarrow a$ ou à sa règle contraire $a \rightarrow \overline{b}$. C'est aussi une mesure qui décroît rapidement dès les premiers contre-exemples. Le taux informationnel *TIC* d'un couple implicatif est nul dès qu'une des deux règles qui le constituent n'apporte pas d'information. Ainsi la mesure permet aussi de repérer à la fois les situations de déséquilibre pour la règle ou sa contraposée et les situations d'indépendance.

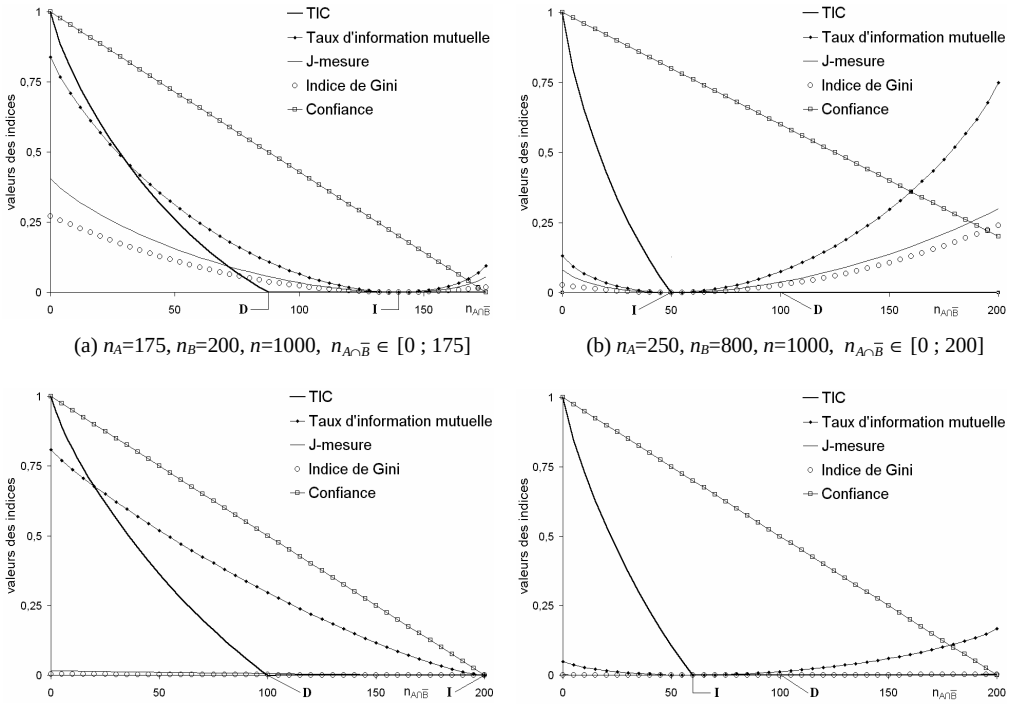
Les figures 3, 4 et 5 permettent de comparer dans différentes configurations de données le taux informationnel *TIC* aux mesures issues de la théorie de l'information habituellement utilisées pour évaluer des règles : le taux d'information mutuelle, la J-mesure, et l'indice de Gini (voir formules dans le tableau 1). Ces figures intègrent également la confiance comme indice de référence.

Dans les figures 3, les mesures sont représentées en fonction du nombre de contre-exemples $n_{A\cap\overline{B}}$. Les figures 3.a et 3.c concernent des configurations de données où le déséquilibre (le premier déséquilibre parmi celui de la règle directe et celui de la règle contraposée, c'est-à-dire $n_{A\cap\overline{B}} = \min(n_A/2, n_{\overline{B}}/2)$) a lieu avant l'indépendance, tandis que les figures 3.b et 3.d concernent des configurations de données où le déséquilibre a lieu après l'indépendance. Ces quatre figures montrent que quelle que soit la configuration, la mesure *TIC* rejette à la fois déséquilibre et indépendance. Le taux d'information mutuelle, la J-

mesure et l'indice de Gini s'annulent à l'indépendance mais ne repèrent pas le déséquilibre (elles peuvent même y prendre des valeurs élevées). La confiance quant à elle vaut 0.5 au déséquilibre de la règle directe ($n_{A\cap\overline{B}} = n_A/2$) mais elle ne repère pas l'indépendance (elle peut même y prendre des valeurs élevées). Il est à noter que seules la confiance et *TIC* permettent de fixer un seuil de filtrage des règles de manière absolue (le maximum vaut toujours 1).

Taux d'information mutuelle	$\left[\sum_i\sum_j P(A_i,B_j)\log(\frac{P(A_i,B_j)}{P(A_i)P(B_j)})\right] / \min(-\sum_i P(A_i)\log(P(A_i)),-\sum_j P(A_j)\log(P(A_j)))$
J-mesure	$P(A,B)\log(\frac{P(B\|A)}{P(B)}) + P(A,\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})})$
Indice de Gini	$P(A)\left(P(B\|A)^2 + P(\overline{B}\|A)^2\right) + P(\overline{A})\left(P(B\|\overline{A})^2 + P(\overline{B}\|\overline{A})^2\right) - P(B)^2 - P(\overline{B})^2$

TAB 1 – *Mesures de qualité de règles issues de la théorie de l'information (formules données pour a → b)*



(a) n_A=175, n_B=200, n=1000, $n_{A\cap\overline{B}} \in [0\ ;\ 175]$

(b) n_A=250, n_B=800, n=1000, $n_{A\cap\overline{B}} \in [0\ ;\ 200]$

(c) n_A=200, n_B=400, n=100 000, $n_{A\cap\overline{B}} \in [0\ ;\ 200]$

(d) n_A=200, n_B=70 000, n=100 000, $n_{A\cap\overline{B}} \in [0\ ;\ 200]$

FIG. 3 – *Evolution des mesures en fonction du nombre de contre-exemples*
(I désigne l'indépendance $n_{A\cap\overline{B}} = n_A \times n_{\overline{B}} / n$ et D le déséquilibre $n_{A\cap\overline{B}} = min(n_A/2, n_{\overline{B}}/2)$)

Les figures 3 illustrent aussi le caractère symétrique du taux d'information mutuelle, de la J-mesure et de l'indice de Gini. L'information mutuelle est invariante par permutation ou négation des variables prémisse et conclusion, puisqu'elle est calculée sur la totalité de la distribution jointe des variables. L'indice de Gini est invariant par négation, et la J-mesure invariante par négation de la conclusion. Tous ces indices informationnels ne respectent donc pas le caractère asymétrique des règles, ce qui justifie l'introduction d'une mesure informationnelle mieux adaptée à la sémantique des règles comme *TIC*.

Les figures 3.c et 3.d sont relatives à des configurations de données où le nombre total d'individus est grand (n = 100 000). Elles illustrent la sensibilité des mesures aux "pépites de connaissance" que sont les règles très spécifiques. Nous pouvons voir que la J-mesure et l'indice de Gini deviennent peu discriminants lorsque n est grand (ils sont quasi-nuls), tandis que la confiance, *TIC* et dans une moindre mesure le taux d'information mutuelle n'occultent pas les cas rares et ont la capacité de repérer les règles spécifiques. Ces comportements sont aussi visibles sur la simulation de la figure 4 dans laquelle n augmente à partir d'une configuration initiale fixe n_A, n_B, $n_{A\cap\overline{B}}$. Pour *TIC* et le taux d'information mutuelle, une règle peu contredite est d'autant meilleure que n est grand, ce qui s'explique par le fait que sa contraposée est de plus en plus confirmée.

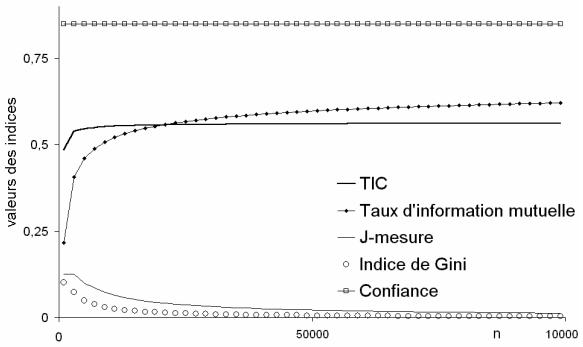
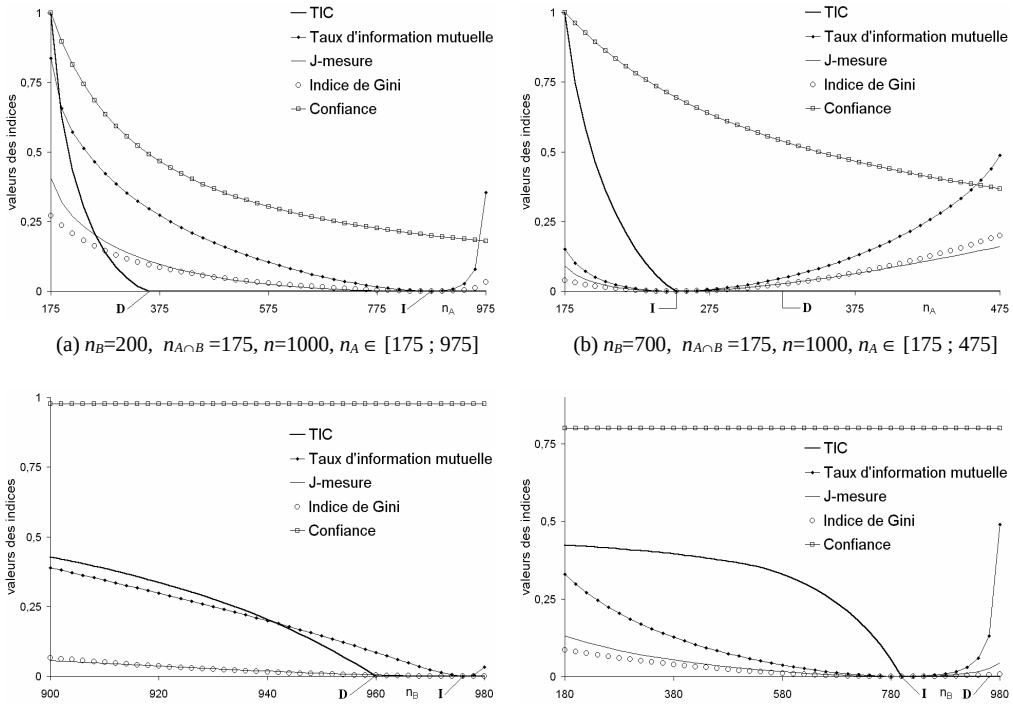


FIG. 4 – *Evolution des mesures avec n_A=200, n_B=400, $n_{A\cap\overline{B}}$ =30, $n \in$ [1000 ; 100 000]*

Dans les figures 5.a-b et 5.c-d, les mesures sont représentées en fonction de n_A et n_B respectivement. Les figures 5.a et 5.b illustrent le comportements des indices lorsque A grandit à l'extérieur de B, tandis que sur les figures 5.c et 5.d, c'est l'ensemble B qui grandit à l'extérieur de A. Comme précédemment, les figures 5.a et 5.c concernent des configurations de données où le déséquilibre a lieu avant l'indépendance, et les figures 5.b et 5.d concernent des configurations de données où le déséquilibre a lieu après l'indépendance. Ces quatre simulations illustrent le comportement satisfaisant de *TIC*, qui décroît lorsque n_A ou n_B augmente. Les trois propriétés caractérisant une bonne mesure de qualité de règles énoncées par Piatetsky-Shapiro [Piatestsky-Shapiro, 1991] sont donc vérifiées par *TIC*. Les figures montrent aussi que *TIC* rejette à la fois déséquilibre et indépendance.



(a) n_B=200, $n_{A\cap B}$ =175, n=1000, $n_A \in$ [175 ; 975] (b) n_B=700, $n_{A\cap B}$ =175, n=1000, $n_A \in$ [175 ; 475]

(c) n_A=850, $n_{A\cap \overline{B}}$ =20, n=1000, $n_B \in$ [900 ; 980] (d) n_A=100, $n_{A\cap \overline{B}}$ =20, n=1000, $n_B \in$ [180 ; 980]

FIG. 5 – *Evolution des mesures quand A ou B grandit*
(I désigne l'indépendance $n_{A\cap \overline{B}} = n_A \times n_{\overline{B}} / n$ *et D le déséquilibre* $n_{A\cap \overline{B}} = min(n_A/2, n_{\overline{B}}/2)$*)*

4. Conclusion

Nous avons présenté dans cet article une nouvelle mesure objective de qualité pour les règles, le *Taux Informationnel modulé par la Contraposée* (*TIC*), qui s'appuie sur la théorie de l'information. Cette mesure est bien adaptée à la sémantique causale des règles, puisque d'une part, contrairement aux autres indices entropiques de qualité de règles, elle respecte leur caractère asymétrique, et d'autre part elle prend en compte la contraposée. A notre connaissance, *TIC* est la seule mesure de qualité de règles qui intègre à la fois indépendance et déséquilibre. Il suffit en effet de ne retenir que les règles informatives pour rejeter simultanément celles établies entre variables corrélées négativement et celles qui possèdent plus de contre-exemples que d'exemples. De plus, les simulations numériques encouragent à utiliser *TIC* pour la recherche de "pépites de connaissance".

La mesure *TIC* évalue la puissance implicative des règles en considérant leur contenu informationnel. Cependant, dans le cadre de la recherche de règles, et même de couples implicatifs, de bonne qualité, il est nécessaire d'évaluer également la significativité statistique des règles. Ainsi, nous comptons coupler *TIC* avec l'intensité d'implication pour l'analyse statistique implicative [Gras, 1996], par exemple au sein d'une mesure synthétique, pour reprendre la paradigme de l'intensité d'implication entropique. Par ailleurs, la mesure

TIC est intégrée au sein de deux de nos outils issus de travaux en cours : l'outil de visualisation *ARVis* pour la fouille interactive de règles [Blanchard *et al.*, 2003b] et la plateforme *ARVal* pour l'évaluation des mesures de qualité [Popovici, 2003].

Références

[Agrawal *et al.*, 1993] R. Agrawal, T. Imielinsky et A. Swami. Mining associations rules between sets of items in large databases. *Proc. of ACM SIGMOD'93*, 1993, p. 207-216

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

[Bayardo et Agrawal, 1999] R.J. Bayardo et R. Agrawal. Mining the most interesting rules. *Proc. of the 5th Int. Conf. on Knowledge Discovery and Data Mining*, 1999, p.145-154

[Blanchard *et al.*, 2003a] J. Blanchard, P. Kuntz, F. Guillet et R. Gras. Implication intensity: from the basic statistical definition to the entropic version. *Statistical Data Mining and Knowledge Discovery*, Bozdogan H. (ed.), CRC Press, 2003, p. 473-485

[Blanchard *et al.*, 2003b] J. Blanchard, F. Guillet et H. Briand. A user-driven and quality-oriented visualization for mining association rules. *Proc. of the 3rd Int. Conf. on Data Mining*, IEEE Computer Society Press, 2003, p. 493-496

[Brin *et al.*, 1997a] S. Brin, R. Motwani et C. Silverstein. Beyond market baskets: generalizing association rules to correlations. *Proc. of ACM SIGMOD'97*, 1997, p. 265-276

[Brin *et al.*, 1997b] S. Brin, R. Motwani, J. Ullman et S. Tsur. Dynamic itemset counting and implication rules for market basket data. *Proc. of the Int. Conf. on Management of Data*, ACM Press, 1997, p. 255-264

[Freitas, 1999] A. Freitas. On rule interestingness measures. *Knowledge-Based Systems Journal 12(5-6)*, 1999, p. 309-315

[Gras, 1996] R. Gras et coll.. *L'implication statistique - Nouvelle méthode exploratoire de données*. La pensée sauvage éditions, 1996

[Gras *et al.*, 2001] R. Gras, P. Kuntz, R. Couturier et F. Guillet. Une version entropique de l'intensité d'implication pour les corpus volumineux. *Extraction des Connaissances et Apprentissage 1(1-2)*, 2001, p. 69-80

[Hilderman et Hamilton, 2001] R. Hilderman et H. Hamilton. *Knowledge discovery and measures of interest*. Kluwer Academic publishers, 2001

[Jaroszewicz et Simovici, 2001] S. Jaroszewicz et D.A. Simovici. A general measure of rule interestingness. *Proc. of the 7th Int. Conf. on Knowledge Discovery and Data Mining*, L.N.C.S. 2168, Springer, 2001, p. 253-265

[Kodratoff, 2001] Y. Kodratoff. Comparing machine learning and knowledge discovery in databases: an application to knowledge discovery in texts. *Machine Learning and Its Applications*, Paliouras G., Karkaletsis V., Spyropoulos C.D. (eds.), L.N.C.S. 2049, Springer, 2001, p. 1-21

[Liu *et al.*, 1999] B. Liu, W. Hsu, L. Mun et H. Lee. Finding interesting patterns using user expectations. *IEEE Transactions on Knowledge and Data Engineering 11*, 1999, p. 817-832

[Loevinger, 1947] J. Loevinger. A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs 61(4)*, 1947

[Padmanabhan et Tuzhilin, 1998] B. Padmanabhan et A. Tuzhilin. A belief-driven method for discovering unexpected patterns. *Proc. of the 4th Int. Conf. on Knowledge Discovery and Data Mining*, 1998, p. 94-100

[Piatestsky-Shapiro, 1991] G. Piatestsky-Shapiro. Discovery, analysis, and presentation of strong rules. *Knowledge Discovery in Databases*. Piatetsky-Shapiro G., Frawley W.J. (eds.), AAAI/MIT Press, 1991, p. 229-248

[Popovici, 2003] E. Popovici. Un atelier pour l'évaluation des indices de qualité. Mémoire de D.E.A. E.C.D., IRIN/Université Lyon2/RACAI Bucarest, Juin 2003

[Sebag et Schoenauer, 1988] M. Sebag et M. Schoenauer. Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases. *Proc. of the European Knowledge Acquisition Workshop (EKAW'88)*, Boose J., Gaines B., Linster M. (eds.), Gesellschaft für Mathematik und Datenverarbeitung mbH, 1988, p. 28.1-28.20

[Shannon et Weaver, 1949] C.E. Shannon et W. Weaver. *The mathematical theory of communication*. University of Illinois Press, 1949

[Smyth et Goodman, 1991] P. Smyth et R.M. Goodman. Rule induction using information theory. *Knowledge Discovery in Databases*, Piatetsky-Shapiro G., Frawley W.J. (eds.), AAAI/MIT Press, 1991, p. 159-176

[Tan *et al.*, 2002] P. Tan, V. Kumar et J. Srivastava. Selecting the right interestingness measure for association patterns. *Proc. of the 8th Int. Conf. on Knowledge Discovery and Data Mining*, 2002, p. 32-41

Summary

Knowledge validation is one of the most problematic steps in an association rule discovery process. To enable the user (a decision-maker specialized in the data studied) to find interesting knowledge in the large amounts of rules produced by the data mining algorithms, it is necessary to measure the quality of the rules. In this article, we propose to assess the rules by considering their informational content with a new interestingness measure based on the Shannon entropy: *TIC* (*informational ratio modulated by the contrapositive*). This index has the advantage of being well suited to the rule semantics, since on the one hand it respects their asymmetric feature and on the other hand it benefits from their contrapositives. Furthermore, to our knowledge it is the only rule interestingness measure which integrates both independence and imbalance, that is to say which allows to reject simultaneously the rules between negatively correlated variables and the rules with more counter-examples than examples. Comparisons of *TIC* with J-measure, mutual information, Gini index, and confidence are realized on numerical simulations.