Analyse du comportement limite d'indices probabilistes pour une sélection discriminante

Sylvie Guillaume* et Israël-César Lerman**

*Clermont Université, Université d'Auvergne, LIMOS, BP 10448, F-63000 Clermont-Fd
sylvie.guillaume@isima.fr
**Irisa, Université de Rennes 1, Campus de Beaulieu, 35042 Rennes Cedex
lerman@irisa.fr

Résumé. Nous étudions ici le comportement de deux types d'indices probabilistes discriminants en présence de données dont le volume va en croissant. À cet égard, un modèle spécifique de croissance de la taille des données et de liaison entre variables est mis en œuvre et celui-ci va permettre de déterminer le comportement limite des différents indices quel que soit le niveau de liaison entre la prémisse et la conclusion de la règle donnée. La clarté des résultats obtenus nous conduit à en chercher l'explication formelle. L'expérimentation a été effectuée avec la base de données UCI *Wages*.

1 Introduction

L'extraction des règles d'association (Agrawal et Srikant 1994) est un domaine largement étudié dans la communauté "*extraction des connaissances*". Une règle d'association est une implication du type $a \rightarrow b$ où a et b sont des conjonctions de variables binaires disjointes. Afin de déterminer les règles intéressantes, deux indices[1] sont communément utilisés : (1) le support $p(a \wedge b)$ (*ou taux de couverture*) qui mesure la fréquence d'apparition de la règle et indique le pourcentage d'individus contenant toutes les variables de la règle, (2) la confiance $p(b/a)$ (*ou probabilité conditionnelle*) qui mesure la force de la règle et évalue le pourcentage d'individus vérifiant la conclusion b parmi ceux qui vérifient la prémisse a. Une règle sera dite valide si les valeurs prises par le couple de mesures sont supérieures à deux seuils fixés par l'utilisateur : le support minimum sup_{min} et la confiance minimum $conf_{min}$. De nombreux auteurs comme par exemple Sese et Morishita (Sese et Morishita 2002) ont montré les faiblesses de ce couple de mesures qui valide des règles qui ne sont pas toujours pertinentes. De nombreux indices ont été proposés dans la littérature pour palier les faiblesses de ces deux mesures. Le lecteur pourra consulter les articles de synthèse (Tan et al. 2002, Lallich et Teytaud 2004, Geng et Hamilton 2007, Feno 2007, Vaillant 2007 et Guillaume et al. 2010) mettant en évidence non seulement les nombreux indices d'intérêt, mais aussi les propriétés de ces mesures d'intérêt, afin d'aider l'utilisateur dans le choix d'une ou plusieurs mesures complémentaires capables d'éliminer les règles valides non pertinentes. Une famille d'indices, celle reposant sur une échelle probabiliste, a montré tout son intérêt puisqu'elle est capable d'éliminer certains types de règles valides inintéressantes. Fleury (Fleury, 1996) et Guillaume (Guillaume, 2000) révèlent les types de règles éliminées. L'indice fondateur de

[1] ou mesures.

cette famille d'indices est l'indice de la vraisemblance du lien (Lerman, 1981), qui évalue la vraisemblance de la "*grandeur*" du nombre d'individus vérifiant à la fois la prémisse *a* et la conclusion *b*. On trouve également l'intensité d'implication (Gras, 1979), qui fonctionne sur le même principe que le précédent, mais qui évalue la "*petitesse*" du nombre de contre-exemples (*i.e. individus qui vérifient la prémisse a mais qui ne vérifient pas la conclusion b*) et non plus l'importance du nombre d'exemples. Ce deuxième indice a l'avantage de donner un sens à l'orientation de l'association entre *a* et *b*, c'est-à-dire de révéler la ou les règle(s) pertinente(s) : $a \rightarrow b$ et/ou $b \rightarrow a$. Cependant cette famille d'indices, bien qu'ayant des vertus éliminatrices de règles inintéressantes, a la particularité de devenir non discriminante en présence de données trop volumineuses : dans ce cas les indices prennent une valeur égale à *1* pour les règles dont la confiance est supérieure à $p(b)$ (*probabilité d'apparition de b*) et une valeur égale à *0* pour les règles dont cette fois-ci la confiance est inférieure à $p(b)$. Afin de remédier à ce problème, des conceptions nouvelles de ces indices ont été élaborées afin que l'échelle de probabilité devienne discriminante pour un nombre important d'individus. Pour cela, une normalisation préalable et indispensable est effectuée. Différentes techniques de normalisation ont été proposées. La première est contextuelle et évalue une règle par rapport à l'ensemble potentiel des règles valides, c'est le cas de l'intensité d'implication contextuelle (Lerman et Azé, 2007) noté *VLgrImpP* (*VLgrImpP pour indice de Vraisemblance du Lien globalement réduit, Implicatif, et se référant à un modèle Poissonnien de l'hypothèse d'absence de liaison*). La seconde raisonne par rapport à un échantillon dont la taille serait réduite à $e = 100$ et propose une Valeur Test (Rakotomalala et Morineau, 2008) pour ce niveau considéré noté *VTeImpBarP* (*VTeImpBarP pour Valeur Test se basant sur un échantillon de taille e = 100, Implicatif avec une approche Barycentrique et se référant à un modèle Poissonnien*). Deux variantes de ce dernier indice ont été proposées dans (Lerman et Guillaume, 2010) : les indices *VTeImpCorP* (*VTeImpCorP pour Valeur Test se basant sur un échantillon de taille e = 100, Implicatif avec une approche Corrélative et se référant à un modèle Poissonnien*) et *VTeImpProj* (*VTeImpProj pour Valeur Test se basant sur un échantillon de taille e = 100, Implicatif avec une approche par Projection sur un ensemble aléatoire*). Enfin, la dernière technique mélange par une opération de moyenne géométrique, un indice probabiliste non normalisé avec un indice d'inclusion faisant appel à l'entropie de Shannon : c'est l'intensité d'implication entropique (Gras et al., 2001). Nous nous limitons ici à l'étude des deux premières catégories d'indices probabilistes discriminants. L'objet de l'article est de comparer leurs comportements respectifs dans un contexte de fouille de données. À cet égard, un modèle spécifique de croissance de la taille des données et de liaison entre variables est mis en œuvre afin d'étudier le comportement limite des différents indices quel que soit le niveau de liaison entre la prémisse et la conclusion de la règle donnée.

L'article s'organise donc de la façon suivante. La *section 2* présente le modèle de variation de la taille des données et de liaison entre variables. La *section 3* se consacre à l'étude du comportement limite des différents indices en présence de données volumineuses et quel que soit le niveau de liaison entre la prémisse et la conclusion de la règle. L'article se termine par une conclusion et des perspectives. En raison d'un manque de place, nous ne rappelons pas dans cet article la définition et la sémantique des différents indices d'implication discriminants *VLgrImpP*, *VTeImpBarP*, *VTeImpCorP* et *VTeImpProj* étudiés mais le lecteur pourra les trouver dans l'article précédent, intitulé "*Comparaison entre deux indices pour l'évaluation probabiliste discriminante des règles d'association*", de ce même volume.

2 Modèle de variation

Le modèle de variation que nous étudions ici, qu'on peut désigner par M_3, est plus général que celui considéré dans (Lerman et Guillaume, 2010) et nommé M_2. Relativement à un couple (a, b) de booléens où a est la prémisse et b la conclusion avec $n(a) < n(b)$ ($n(a)$ *étant le nombre d'individus vérifiant la prémisse* a), considérons les ensembles $\mathscr{D}$ (*ensemble des unités de données*), A (*ensemble des individus vérifiant la prémisse* a), B et $A \cap B$. Alors que pour M_2, les ensembles A, B et $A \cap B$ sont fixés et que seul $\mathscr{D} - A \cup B$ croît à partir de sa valeur initiale (*voir l'axe vertical de la figure 1*), ici A est fixe pour $\mathscr{D}$ donné et B garde son cardinal $n(b)$. L'ensemble B commence par être inclus dans $\mathscr{D} - A$, puis se déplace graduellement en intégrant à chaque pas un élément de A et en abandonnant un élément dans $\mathscr{D} - A \cup B$ et ce jusqu'à inclure totalement A (*voir l'axe horizontal de la figure 1*). Ainsi, on passe de l'*incompatibilité* ($n(a \wedge b) = 0$) entre a et b jusqu'à l'*implication logique* ($n(a \wedge b) = n(a)$). Un protocole expérimental a été conçu permettant de réaliser algorithmiquement le modèle M_3.

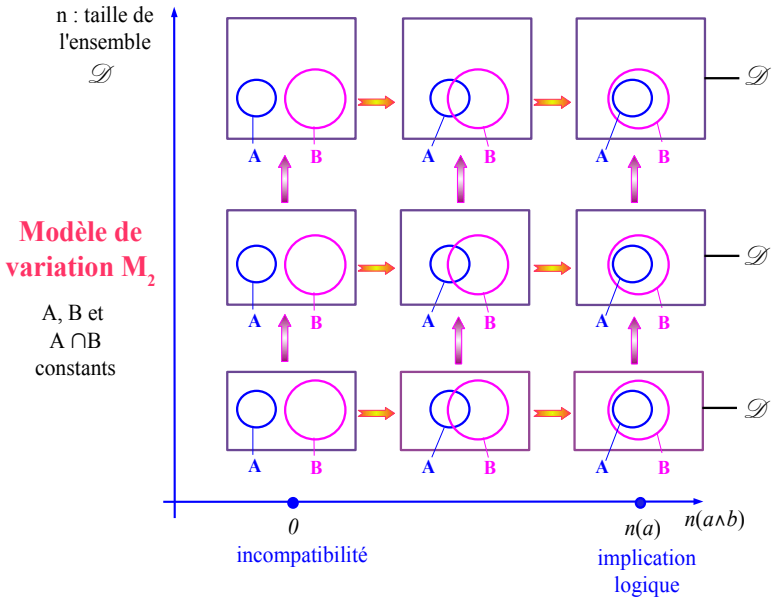

FIG. 1 – *Modèle de variation* M_3.

Le protocole expérimental pour le modèle de croissance exposé précédemment diffère selon les indices. En effet, comme l'indice *VLgrImpP* est une mesure contextuelle, celui-ci a besoin d'un contexte d'extraction et donc d'un ensemble de données dans lequel l'ensemble des règles valides doit être extrait contrairement aux trois autres indices. Nous commençons par décrire le protocole expérimental pour les trois indices de la famille *VTe* et nous terminons par celui de *VLgrImpP*.

Le premier protocole expérimental est simple et nécessite la connaissance des paramètres suivants : $n(a)$, $n(b)$ et l'ensemble $N = \{n_1, n_2, .., n_i, .., n_p\}$ des valeurs que peut prendre la

taille des données. Ce protocole est le suivant : pour chaque valeur n_i de l'ensemble N et pour chaque valeur $n(a \wedge b)$ possible comprise entre 0 et $n(a)$, calculer les valeurs des trois indices de la famille *VTe* : *VTeImpBarP*, *VTeImpCorP* et *VTeImpProj*. Une fois toutes les valeurs obtenues, nous traçons les courbes d'évolution des différents indices (*valeurs des indices en ordonnées pour les courbes des figures 2 et 3*) en fonction du nombre d'exemples $n(a \wedge b)$ (*axe des abscisses pour les courbes des figures 2 et 3*) et pour différentes valeurs de n.

Le deuxième protocole expérimental pour l'indice *VLgrImpP* fait intervenir une base de données et le choix arbitraire d'une règle $a \rightarrow b$ tel que $n(a) < n(b)$. La seule difficulté vient du fait qu'il faut faire varier le nombre d'exemples $n(a \wedge b)$ car pour une règle donnée, ce nombre observé d'exemples est fixe dans la base de données.

Après avoir exposé le modèle de variation, nous étudions maintenant le comportement limite des différents indices.

3 Comportement limite des différents indices

Pour l'étude comportementale de l'indice *VLgrImpP*, nous avons utilisé à titre d'illustration la base de données "*Wages*", données disponibles sur UCI KDD archive (Frank et Asuncion, 2010), et choisi à titre d'illustration la règle "*syndiqué* → *féminin*" pour laquelle $n(a) = 96$, $n(b) = 245$, $n(a \wedge b) = 28$ et $n = 534$. L'ensemble N retenu est le suivant {*534*, *800*, *1 200*, *2 000*, *10 000*, *60 000*, *100 000*}, ce qui nous a obligé, en raison des fortes valeurs de n, à retenir un seuil minimal pour le support égal à 0 ($min_{sup} = 0$). Nous avons retenu le même seuil pour la confiance ($min_{conf} = 0$) afin d'avoir un nombre important de règles mais un autre choix pour ce seuil était tout à fait possible.

En raison du choix de la règle "*syndiqué* → *féminin*", nous avons retenu les paramètres suivants pour tracer les courbes d'évolution des indices de la famille *VTe* : $n(a) = 96$ et $n(b) = 245$. Nous avons également retenu le même ensemble N que pour l'indice *VLgrImpP*.

Les courbes des *figures 2* et *3* nous restituent l'évolution des différents indices pour le modèle de croissance exposé précédemment (*axe des abscisses* : $n(a \wedge b)$, *axe des ordonnées* : *valeurs des indices étudiés*).

Nous constatons que les quatre indices ont des courbes similaires pour la valeur initiale $n = 534$: une courbe en S avec des valeurs comprises entre 0 et 1, et une pente plus importante pour *VLgrImpP*. Ensuite les courbes des indices de la famille *VTe* sont également similaires jusqu'à la valeur $n = 2\,000$ avec des valeurs pour les indices comprises cette fois-ci entre $0{,}3$ et $0{,}98$. Au delà de la valeur $n = 2\,000$, les courbes en S se transforment en droites. Pour l'indice *VTeImpBarP*, plus la valeur de n est élevée et plus ces courbes tendent vers des droites parallèles à l'axe des x et avec des valeurs pour la mesure inférieure à $0{,}1$. Même constat pour l'indice *VTeImpCorP* mais les droites sont légèrement plus pentues que celles de la mesure *VTeImpBarP*. Quant à la mesure *VTeImpProj*, les droites sont translatées vers des valeurs plus élevées pour la mesure c'est-à-dire des valeurs comprises entre $0{,}5$ et $0{,}65$.

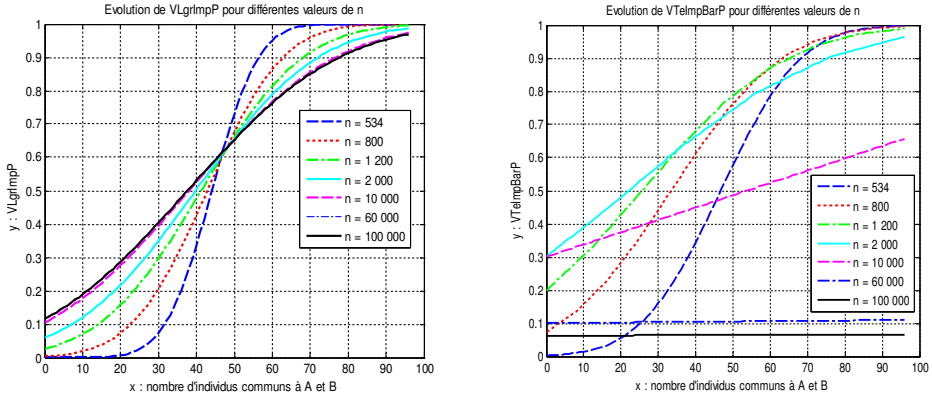

FIG. 2 – *Évolution des indices VLgrImpP (courbe de gauche) et VTeImpBarP (courbe de droite) pour différentes valeurs de n.*

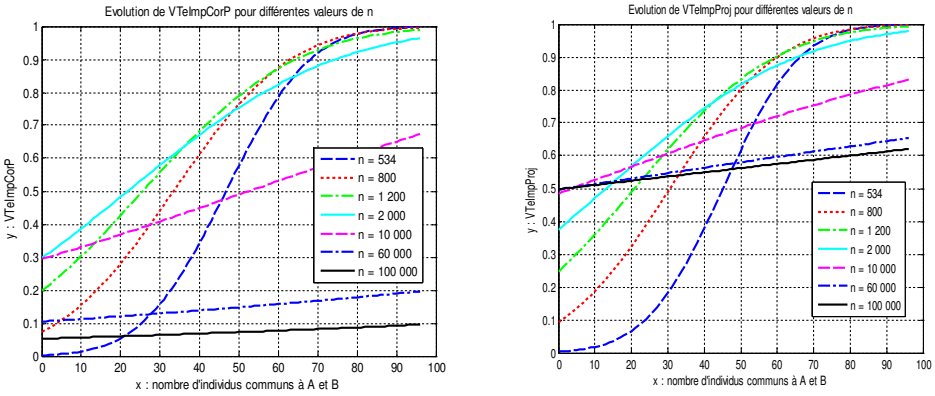

FIG. 3 – *Évolution des indices VTeImpCorP (courbe de gauche) et VTeImpProj (courbe de droite) pour différentes valeurs de n.*

Pour finir, l'indice *VLgrImpP* garde pour les différentes valeurs de *n* des courbes en *S* et ces courbes tendent très rapidement vers une courbe limite qui commence à se dessiner pour la valeur *n = 10 000*. Cette courbe limite prend des valeurs comprises entre les valeurs *0,12* et *0,98*, donc une amplitude (*0,98 - 0,12 = 0,86*) assez élevée comparativement aux trois autres indices (*amplitude quasiment nulle pour VTeImpBarP, amplitude de 0,05 pour VTeImpCorP et une amplitude de 0,12 pour VTeImpProj*).

Les courbes obtenues pour les indices de la famille des *VTe* ne sont pas surprenantes car dans (Lerman et Guillaume, 2010), il a été démontré dans la *section 5.3.3 page 29* que lorsque la taille *n* de l'ensemble des données augmente, la courbe d'évolution des trois indices pour une règle donnée *a* → *b* est d'abord croissante puis décroissante. La *figure 4* nous restitue cette croissance puis cette décroissance pour les trois indices pour la règle précédemment retenue : "*syndiqué* → *féminin*" (*nous rappelons que le nombre d'exemples pour cette règle est de 28*).

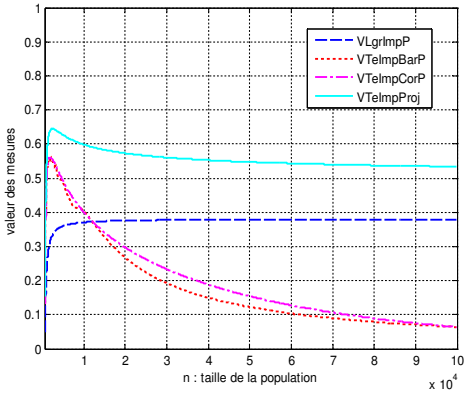

FIG. 4 – *Évolution des différents indices pour la règle "syndiqué → féminin" lorsque la taille de l'ensemble des données augmente jusqu'à 100 000 et dans le cas où n(a∧b) = 28.*

Nous vérifions l'évolution des courbes des *figures 2* et *3* et constatons une valeur plus élevée pour l'indice *VTeImpProj* puisque celle-ci est d'environ *0,54* pour *n=100 000* et *n*(*a*∧*b*) = *28*, suivie par l'indice *VLgrImpP* puisque la valeur est d'environ *0,38*. Quant aux indices *VTeImpBarP* et *VTeImpCorP* les valeurs sont très faibles et sont de l'ordre de *0,08*. Dans cette situation (*cas où n*(*a*∧*b*) = *28*), nous vérifions l'insensibilité de *VLgrImpP* à la taille de *n* au delà d'une certaine valeur qui est ici d'environ *10 000*, ce qui confirme l'apparition de la courbe limite pour l'indice *VLgrImpP* (*courbe de gauche de la figure 2*). Cette situation pour l'indice *VLgrImpP* où à partir d'une certaine valeur pour *n*, la valeur de l'indice devient invariante a été mise en évidence dans (Lerman et Guillaume, 2010) *figure 36*, *page 56* et ceci quel que soit le niveau d'implication entre la prémisse et la conclusion. Cette constance dans les valeurs de l'indice *VLgrImpP* explique l'apparition de cette courbe limite en *S*.

Les courbes limites mises en évidence pour les indices de la famille *VTe* (*droites faiblement pentues*) nous indique que dans le cas de données volumineuses, ces indices restituent des valeurs très proches pour les différents niveaux d'implication entre la prémisse et la conclusion. Ces indices ont donc des difficultés pour différencier des règles proches de l'incompatibilité avec des règles proches de l'implication logique. De plus, et ceci est particulièrement vrai pour les indices *VTeImpBarP* et *VTeImpCorP*, le problème de la détermination d'un seuil au delà duquel les règles sont pertinentes se pose.

Cette étude nous a donc révélé qu'en présence de données volumineuses, l'indice *VLgrImpP* est la mesure la plus apte à différencier les différents niveaux d'implication entre la prémisse et la conclusion et qu'au delà d'une certaine valeur pour *n*, cet indice devient insensible au volume très important des données. Cette stabilité limite peut se comprendre compte tenu du caractère relatif de *VLgrImpP* par rapport à l'ensemble $\mathscr{R} = \{(a,b) \in \mathscr{A} \times \mathscr{A} \ / \ p(b/a) \geq min_{conf} \text{ et } p(a \wedge b) \geq min_{sup} \}$ des règles valides (avec $\mathscr{A}$ *l'ensemble des attributs binaires*). On se rend compte dans (Lerman et Guillaume 2010) *sous-section 5.3.3* que pour un couple d'attributs donné (*a*,*b*), *a*→*b*, lorsque *n* augmente sous le modèle M_2, la variation relative de l'indice brut centré réduit $Q(a \wedge \bar{b}) = \frac{n(a \wedge \bar{b}) - n(a) \times n(\bar{b})/n}{\sqrt{n(a) \times n(\bar{b})/n}}$ diminue ($n(a \wedge \bar{b})$ *est le nombre d'individus vérifiant*

a et ne vérifiant pas b). Cette dernière propriété étant vraie pour tous les couples d'attributs faisant partie de $\mathscr{R}$, on a bien de façon relative la stabilité du comportement limite de $Q(a \wedge \bar{b})$ et donc de $VLgrImpP(a \rightarrow b) = 1 - \phi(Q(a \wedge \bar{b}))$ (ϕ *étant la fonction de répartition de la loi normale centrée réduite*).

4 Conclusion et perspectives

Dans cet article, nous avons étudié deux catégories d'indices probabilistes discriminants : la première catégorie est contextuelle et évalue une règle par rapport à l'ensemble potentiel des règles valides, c'est l'indice *VLgrImpP* ; la seconde catégorie raisonne par rapport à un échantillon dont la taille serait réduite à une valeur *e* (*en général e =100*) et utilise une Valeur Test noté *VTe* pour ce niveau considéré, ce sont les indices *VTeImpBarP*, *VTeImpCorP* et *VTeImpProj*. Ces deux familles d'indices ont été mises en oeuvre pour répondre au problème de la non discrimination des indices implicatifs probabilistes dans un contexte de fouille de données. L'objectif de cet article était d'une part, de savoir si ces indices étaient réellement discriminants en présence de données volumineuses et d'autre part, s'ils étaient équivalents. Pour cela, nous avons effectué une série d'expériences qui a permis de détecter le comportement limite de ces indices en présence de données volumineuses et ceci quel que soit le niveau de liaison entre la prémisse et la conclusion de la règle. Il est apparu à l'issue de cette série d'expériences que l'indice *VLgrImpP* est le plus performant pour deux raisons. Tout d'abord, l'indice devient insensible à la taille des données à partir d'une certaine valeur de celle-ci, et restitue des valeurs quasiment identiques pour un niveau de liaison donné entre la prémisse et la conclusion. Ensuite, il est capable de différencier des règles avec des niveaux de liaison entre la prémisse et la conclusion relativement proches, contrairement à la seconde famille d'indices pour qui cela devient très difficile.

Après cette étude sur ces deux familles d'indices d'implication probabilistes discriminants, il serait intéressant d'étudier le comportement d'autres indices et notamment de la troisième famille : la famille des intensités d'implication entropiques.

Références

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th VLDB Conference*, Santiago, Chile.

Feno, D.J. (2007). *Mesures de qualité des règles d'association : normalisation et caractérisation des bases.* PhD thesis, Université de La Réunion.

Fleury, L. (1996). *Extraction de Connaissances dans une Base de Données pour la Gestion des Ressources Humaines, Mesure de la qualité d'une règle, Élimination de la redondance, Proposition d'Algorithmes.* Thèse d'État, Nantes, Novembre 1996.

Frank, A. et A. Asuncion (2010). UCI Machine Learning Repository. [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Geng, L. et H.J. Hamilton (2007). Choosing the Right Lens: Finding What is Interesting in Data Mining. In Quality measures in data mining 2007, Volume 43 of *Studies in Computational Intelligence*, pp. 3–24, Springer, ISBN 978-3-540-44911-9.

Gras, R. (1979). *Contribution à l'Etude Expérimentale et à l'Analyse de certaines Acquisitions Cognitives et de certains Objectifs Didactiques en Mathématiques*. Thèse d'Etat, Université de Rennes I, Octobre 1979.

Gras, R., P. Kuntz, R. Couturier et F. Guillet (2001). Une version entropique de l'intensité d'implication pour les corpus volumineux. *Extraction des connaissances et apprentissage* (*EGC 2001*), 1(1-2), pp. 69–80.

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

Guillaume S., D. Grissa et E. Mephu Nguifo (2010). Propriétés des mesures d'intérêt pour l'extraction des règles. In *Actes de l'atelier "Qualité des Données et des Connaissances" (QDC 2010) de la conférence "Extraction et Gestion des Connaissances"*, janvier 2010, pp. 15-28, Hammamet, Tunisie.

Lallich, S. et O. Teytaud (2004). Evaluation et validation de mesures d'intérêt des règles d'association. In *Mesures de Qualité pour la Fouille de Données 2004*, Volume RNTI-E-1 of *RNTI*, pp. 193-217. Cépaduès.

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

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, ISBN 978-3-540-44911-9.

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.

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

Sese, J. et S. Morishita (2002). Answering the most correlated n association rules efficiently. In *Proceedings of the 6th European Conf. on Principles of Data Mining and Knowledge Discovery*, pp. 410–422. Springer-Verlag.

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

Vaillant, B. (2007). *Mesurer la qualité des règles d'association : études formelles et expérimentales*, PhD thesis, ENST Bretagne.

Summary

In this paper, discriminant probabilistic interestingness association rule measures are compared. This comparison is based on a specific simulation model which increases the data size. The limit form of these measures for different association levels is revealed. The well known UCI "*Wages*" data base is employed for experimentation.