Étude de l'interaction entre variables pour l'extraction des règles d'influence

L. Nemmiche Alachaher* et S. Guillaume*

*LIMOS, UBP UMR 6158 CNRS
Complexe des Cézeaux
63177 AUBIERE Cedex - France
{nemmiche, sylvie.guillaume}@isima.fr

Résumé. Cet article présente une méthode efficace pour l'extraction de règles d'influence quantitatives positives et négatives. Ces règles d'influence introduisent une nouvelle sémantique qui vise à faciliter l'analyse d'un volume important de données. Cette sémantique fixe la direction de la règle entre deux variables en positionnant, au préalable, l'une comme étant l'*influent* et l'autre comme étant l'*influé*. Elle permet, de ce fait, d'exprimer la nature de l'influence : *positive*, en maximisant le nombre d'éléments en commun ou *négative*, en maximisant le nombre d'éléments qui violent l'influé.
Notre approche s'appuie sur une stratégie qui comporte cinq étapes dont deux exécutées en parallèle. Ces deux étapes constituent les étapes clé de notre approche. La première combine une méthode d'élagage et de regroupement tabulaire basée sur les tableaux de contingence. Cette dernière construit et classe les zones potentiellement intéressantes. La seconde, injecte la sémantique et évalue le degré d'influence que produirait l'introduction d'une nouvelle variable sur un ensemble de variables en utilisant une nouvelle mesure d'intérêt, l'*Influence*. Cette étape vient affiner les résultats de la première étape, et permet de se focaliser sur des zones valides par rapport aux contraintes spécifiées. Enfin, un système de règles d'influence jugées intéressantes est construit basé sur la juxtaposition des résultats des deux étapes clé de notre approche.

1 Introduction

L'extraction de connaissances est un processus qui permet d'analyser une masses de données importante afin d'en extraire des connaissances nouvelles, valides et utiles. Ces connaissances sont ensuite présentées sous différentes formes notamment sous forme de règles d'association. Une règle d'association (RA) (Agrawal et al. (1993)) est une implication de la forme $C_1 \rightarrow C_2$, où C_1 et C_2 sont des conditions C sur les attributs de la base. Soient $minsup$ et $minconf$ des seuils prédéfinis. Une RA est dite forte si elle satisfait deux contraintes :

- son support $supp(C) \geq minsup$, avec $supp(C)$: nombre de transactions dans la base qui satisfont l'ensemble des conditions C tel que $supp(C_1 \rightarrow C_2) = supp(C_1 \wedge C_2)$;
- sa confiance $conf(C_1 \rightarrow C_2) \geq minconf$, avec $conf(C_1 \rightarrow C_2) = \frac{supp(C_1 \rightarrow C_2)}{supp(C_1)}$.

Dans cet article, nous nous intéressons tout particulièrement à l'extraction de règles d'association quantitatives (RAQ). Ce type de règles prend en considération tout type de variables, quantitatives ou catégorielles. Un certain nombre de travaux relatifs à l'extraction de RAQ sont proposés dans (Srikant et Agrawal (1996), Hong et al. (1999), Aumann et Lindell (1999), Miller et Yang (1997), Lent et al. (1997), Fukuda et al. (1996), Rastogi et Shim (1999), Mata et al. (2002)). L'interprétation des résultats obtenus est basée sur la sémantique choisie. Cette sémantique exprime la nature des liens existants entre les différentes variables de la base. On peut alors exprimer le lien qui maximise le nombre d'éléments en commun entre deux variables. Ce type d'extraction est connu sous le nom d'extraction de règles d'association positives : *une personne qui achète du lait achète aussi du café*. De la même façon, on peut exprimer le lien qui unit deux variables et qui prend en considération le changement de comportement d'une variable spécifique lors de l'introduction d'une nouvelle variable. Cet aspect est particulièrement subtil dans le sens où il considère les variables qui ont un comportement opposé lors de leur union. Les règles extraites à partir de ce type d'union sont connus sous le nom de règles d'association négatives : *une personne qui achète du Pepsi n'achète pas de Coca Cola*,(Nemmiche et Guillaume (2006); Teng et al. (2002); Antonie et Zaane (2004); Wu et al. (2004); Savasere et al. (1998); Yuan et al. (2002); Yan et al. (2004)).
Une règle d'association négative (RAN) est présentée sous l'une des formes suivantes : $C_1 \rightarrow \neg C_2$, $\neg C_1 \rightarrow C_2$ ou $\neg C_1 \rightarrow \neg C_2$.
Dans (Teng et al. (2002)), la distance du χ^2 permet d'évaluer l'indépendance des variables mais sans donner aucune indication sur la dépendance. Dans (Antonie et Zaane (2004)), la mesure d'inégalité de Cauchy Schwarz traduit, plus ou moins, la dépendance linéaire entre deux variables, mais reste cependant incapable de juger de la pertinence réelle de la règle d'association. De plus, elle ne permet pas d'exprimer l'effet d'influence recherché dans notre travail.
$\grave{A}$ travers cet article, nous présentons une nouvelle méthode d'extraction de règles d'association quantitatives positives et négatives. Les règles visées ont une sémantique particulière qui permet, par ailleurs, de faire ressortir l'effet qu'une variable nommée *influent* pourrait avoir sur une autre nommée *influé*. Cet effet est exprimé par un changement significatif du comportement de l'influé. Plus précisément, par une chute importante du nombre d'individus dans l'intersection de l'association étudiée. L'évaluation du degré de changement de comportement est réalisée grâce à une nouvelle mesure nommée *Influence*. Cette évaluation vient affiner les résultats d'une autre étape exécutée en parallèle et qui par l'utilisation des tableaux de contingence permet une sélection et un regroupement tabulaire de zones qui constituent la source de règles potentiellement intéressantes.
Grâce aux filtrages réalisés dans ces deux étapes, des règles que nous avons nommées *règles d'influence* sont extraites basées sur des contraintes spécifiées.
Cet article s'organise de la façon suivante. Dans la Section 2, nous faisons état de quelques travaux sur l'extraction de RAQ positives et négatives. La section 3 expose notre stratégie de génération de règles d'influence. Enfin, nous présentons nos conclusions dans la section 4.

2 Les règles d'association quantitatives positives et négatives

Les RAQ positives prennent en considération les variables qui apparaissent simultanément. Intéressons nous à l'exemple suivant adapté de (Brin et al. (1997)) où A et B sont des ensembles de motifs. Soient $minconf = 0{,}80$ et $minsup = 0{,}40$ et soient $supp(A) = 0{,}25$

et $supp(A \cup B) = 0{,}2$. La règle ($A \rightarrow B$) a un $supp(A \rightarrow B) = 0{,}2$ et une $conf(A \rightarrow B) = \frac{supp(A \cup B)}{supp(A)} = 0{,}8$. Dans ce cas, la règle ($A \rightarrow B$) est considérée comme étant une règle $forte$. Supposons que $supp(B) = 0{,}6$, $supp(A) = 0{,}4$ et $supp(A \cup B) = 0{,}05$, avec $minconf = 0{,}52$. Dans ce cas, $conf(A \rightarrow B) = \frac{supp(A \cup B)}{supp(A)} = 0{,}125 < minconf$ et $supp(A \cup B) = 0{,}05 << 0{,}4$. Dans ce cas, $(A \cup B)$ est un motif non fréquent et la règle $A \rightarrow B$ ne peut être extraite de la base de données vu le support et la confiance fixés.
Cependant, $supp(A \rightarrow \neg B) = supp(A) - supp(A \cup B) = 0{,}35$ et $conf(A \rightarrow \neg B) = \frac{supp(A \cup \neg B)}{supp(A)} = 0{,}875 > minconf$. Dans ce cas la règle $A \rightarrow \neg B$ serait valide pour la base de données.
Toutefois, quelle que soit la nature de la RA, positive ou négative, le support et la confiance utilisés comme uniques contraintes sont insuffisants pour une extraction optimale des RA. Afin de remédier à l'extraction prohibitive des RAQ, plusieurs mesures évaluant la qualité des règles ont été proposées ($Lift$ IBM (1996), $Conviction$ Brin et al. (1997)).
Dans ce papier, nous élargissons l'extraction à ces règles qui sont complémentaires aux règles positives. Dans (Wu et al. (2004)), les RAQ négatives sont extraites à partir d'ensembles de motifs non fréquents. Leur objectif est d'identifier les caractères qui peuvent être ignorés.
Dans (Antonie et Zaane (2004)), une nouvelle mesure statistique appelée l'*inégalité de Cauchy Schwarz* symbolisée par ρ est utilisée à des fins d'organisation dans le rayonnage des supermarchés. Soient A et B deux variables, $\rho = \frac{Cov(A,B)}{\delta_A \delta_B} \in [1, -1]$, avec $Cov(A,B)$la covariance des variables A et B et δ_A et δ_B leurs déviations standards. Sachant que

$$\rho = \begin{cases} +1(corrélation \ \ positive) \\ -1(corrélation \ \ négative) \end{cases} \quad (1)$$

Un coefficient ≈ 0 signifie l'absence de relation linéaire entre A et B mais ne donne aucune indication sur l'indépendance.
Si le coefficient $\approx +1$ ou ≈ -1, A et B sont fortement corrélées. Ceci dit, la corrélation ne doit aucunement être confondue avec la causalité. La corrélation entre deux caractères n'implique absolument pas que l'un cause l'autre.
Des problèmes similaires sont apparus avec l'utilisation de certaines mesures statistiques. La distance du χ_2 (Teng et al. (2002)) détecte les variables qui expriment de fortes liaisons. Elle permet de déterminer si deux variables sont indépendantes. Malheureusement, elle est incapable de donner la direction de la dépendance : $A \rightarrow B$ ou $B \rightarrow A$?
La section suivante introduit une nouvelle sémantique proche des règles d'impacte proposées par Webb dans (Webb (2001)).

3 Extraction des règles d'influence

La présente section introduit une nouvelle sémantique de règles (Nemmiche et Guillaume (2006)) de la forme $influent(s) \rightarrow influé(s)$. L'influent est formé d'une disjonction de différentes valeurs de variables et l'influé est sous forme de conjonction de disjonctions de valeurs de variables.
Dans ce qui suit nous présentons notre processus d'extraction de règles d'influence (Nemmiche et Guillaume (2006)) qui est composé de cinq étapes :

1. préparation de données,

2. évaluation du taux d'influence,
3. détermination des zones intéressantes,
4. coordination des résultats,
5. extraction des règles d'influence.

3.1 Préparation des données

Les données présentes dans les bases de données sont sous forme brute. Antérieurement à leur exploitation, il est nécessaire de les traiter. Ce traitement, dans le cas des variables qualitatives, consiste à éclater une variable en un ensemble de variables correspondantes aux différentes valeurs distinctes de la variable traitée.
Dans le cas des variables quantitatives, le traitement des données est plus complexe, et émane des nouveaux problèmes principalement liés au nombre important de valeurs distinctes que peut prendre une seule variable quantitative. La discrétisation fût une réponse naturelle aux problèmes rencontrés (Srikant et Agrawal (1996); Aumann et Lindell (1999)).
Notre travail de préparation des données consiste à réaliser un codage disjonctif complet précédé d'une étape de discrétisation sur les variables quantitatives. La discrétisation est réalisée en utilisant une méthode de clustering basée sur le calcul des plus proches voisins. Les clusters sont constitués d'individus dont les distances des uns par rapport aux autres n'excèdent pas un seuil limite prédéfini le $MaxDif$.
L'exemple suivant est extrait de la base de données $Wages$[1]. Les clusters de la variable *âge* sont calculés avec $MaxDif = 5$ ans et sont listés dans la table 1.

cluster	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
effectifs	61	105	115	84	61	42	38	28
min	18	24	30	36	42	48	54	60
max	23	29	35	41	47	53	59	64

TAB. 1 – *Exemple de discrétisation de la variable quantitative 'âge'*

3.2 Évaluation des taux d'influence

Afin d'évaluer l'influence d'une variable sur une autre, le rejet de l'hypothèse d'indépendance doit être prouvé. Ceci peut être réalisé grâce à différentes mesures (χ^2, coefficient de corrélation linéaire ...). Suite à cela, une analyse de la nature de la dépendance est réalisée grâce à notre mesure l'$Influence$.
L'$Influence$ ($Infl$) est une mesure qui évalue le nombre d'individus dans $A \backslash B$, c-à-d, l'ensemble des individus dans A qui violent B. Les résultats nous permettent de déterminer si l'influence est de nature positive (*similarité*) symbolisée par $Infl^+$ ou négative (*dissimilarité*) symbolisée par $Infl^-$.

[1] $Wages$: 534 transactions et 11 variables (tirée de la base UCI Murphy et Aha (1995)).

3.2.1 Mesure d'influence

l'$Infl^-$ maximise l'ensemble des individus dans $A\backslash B$ et de ce fait, minimise l'ensemble des individus dans $A \cap B$, (voir figure 1 : i). Les règles d'influence attendus (RI) sont de la forme : $A_i \rightarrow \neg B_j$ où A_i est la $i$éme valeur de A et correspond à l'influent et B_j est la $j$éme valeur de B et correspond à l'Influé.
L'$Infl^-$ de l'association (A_i,B_j) est calculée grâce à l'équation (2).

$$Infl^-{}_{A_i/B_j} = \frac{[|B_j| - |B_j \cap A_i|]}{|B_j|} = Pr(\neg A/B) \quad (2)$$

où $|X|$ symbolise la cardinalité de l'ensemble des individus qui vérifient la condition X.
Les résultats obtenus sont filtrés afin de ne retenir que les taux d'influence supérieur au seuil d'influence négative minimum fixé, symbolisé par S^-, tel que :

- $Infl^-$ $(A \rightarrow \neg B) >= S^- \Rightarrow$ la règle $(A \rightarrow \neg B)$ est potentiellement intéressante.
- $Infl^-$ $(A \rightarrow \neg B) < S^- \Rightarrow$ la règle $(A \rightarrow \neg B)$ n'est pas intéressante.

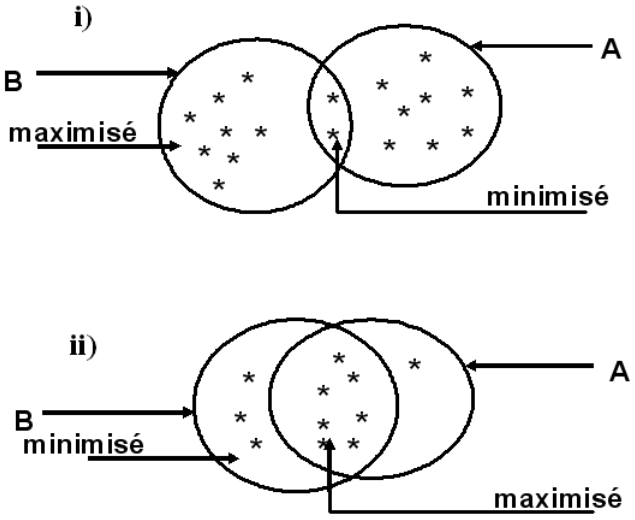

FIG. 1 – *i)Influence$^-$; ii)Influence$^+$*

De la même façon, l'$Infl^+$, maximise le nombre d'individus dans $A \cap B$, minimisant ainsi l'ensemble des individus dans $A\backslash B$, (voir figure 1 : ii). Les règles d'influence attendus sont de la forme $A_i \rightarrow B_j$. Avec, $Infl^+$ de l'association $(A_i,B_j) = 1 - Infl^-$. L'$Infl^+$ de l'association (A_i,B_j) est calculée grâce à l'équation (3).

$$Infl^+{}_{A_i/B_j} = 1 - \frac{[|B_j| - |B_j \cap A_i|]}{|B_j|} = Pr(A/B) \quad (3)$$

Les résultants obtenus, comme dans le cas négatif, sont filtrés par rapport à un seuil d'influence positive minimum fixé et symbolisé par S^+, tel que :

– $Infl^{+}$ $(A \rightarrow B) >= S^{+} \Rightarrow$ la règle $(A \rightarrow B)$ est potentiellement intéressante.
– $Infl^{+}$ $(A \rightarrow B) < S^{+} \Rightarrow$ la règle $(A \rightarrow B)$ n'est pas intéressante.

3.2.2 Consolidation de la mesure d'influence

La dépendance entre deux variables doit être évaluée objectivement. Les ensembles avec peu de représentants sont exclus pour n'inclure que ceux avec un taux d'influence excédant le seuil minimum fixé. Or, si nous fixons le seuil à S=70%, par exemple, alors que le taux d'influence calculé d'une règle est égal à 69%, *ne devrions nous pas prendre cette dernière en considération ?*
Dans notre approche, nous nous sommes intéressées à la pertinence de ce genre de règles et au moyen de les inclure dans notre système. Nous avons, donc utilisé le théorème de la limite centrale (Macfie et Nufrio (2005)) qui répond parfaitement au problème moyennant des intervalles de confiance et une estimation du risque α pour la proportion. Cette évaluation dote la mesure d'influence d'une meilleure sensibilité en permettant une étude dans des intervalles de la forme $[Infl^{-} - \Delta; Infl^{-} + \Delta]$ où Δ est évalué avec l'équation (4) et de la forme $[Infl^{+} - \Delta'; Infl^{+} + \Delta']$ où Δ' est évalué avec l'équation (5). Δ et Δ' sont contraints par le nombre d'éléments qui doit être (≥ 5).

$$\Delta = \beta \times \sqrt{\frac{Infl^{-} \times (1 - Infl^{-})}{|n|}} \tag{4}$$

$$\Delta' = \beta \times \sqrt{\frac{Infl^{+} \times (1 - Infl^{+})}{|n|}} \tag{5}$$

avec α suivant la loi normale $N(0;1)$, β suivant la loi normale inverse et $|n|$ la taille de la base de données.
L'exemple qui suit est tiré de la base $Wages$. Cet exemple permet d'évaluer l'influence négative de la variable $sexe$ représentée par deux clusters (masculin et féminin) sur la variable $salaire$ qui est représentée par sept clusters listés dans la table (2), avec un risque $\alpha = 0{,}05$, $\beta_{\alpha/2} = 1.96$ et $S^{-} = 0{,}70$.

clusters (Salaire)	effectifs	min	max
C1	186	1,0000	6,0000
C2	205	6,1000	11,0200
C3	97	11,1100	16,0000
C4	25	16,1400	20,5500
C5	19	21,2500	26,0000
C6	1	26,2900	26,2900
C7	1	44,5000	44,5000

TAB. 2 – *Discrétisation de la variable quantitative 'salaire'*

Cet exemple fait ressortir trois RI potentiellement intéressantes, (voir les cellules grisées de la table (3), En se référant aux contraintes spécifiées, $|Influé| \geq 5$ et $Infl^{-} \geq S^{-}$ avec S^{-} élément de l'intervalle de confiance. On peut noter dans la table (3), que la ligne en gras

représente une règle avec une $Infl^{-} = 0{,}66$ et grâce au théorème de la limite centrale, ce taux d'influence est considéré comme valide.
Les résultats obtenus lors de cette étape sont ensuite juxtaposés aux résultats de l'étape de détermination des zones intéressantes pour validation.

	$\|sexe\|$	$\|salaire\|$	$\|sexe \cap salaire\|$	$Influence^{-}_{sexe}$	intervalle conf.
M	289	186	79	0,58	[0,53 ;0,62]
a	289	205	110	0,46	[0,42 ;0,51]
s	289	97	64	0,34	[0,30 ;0,38]
c	289	25	21	0,16	[0,13 ;0,19]
u	289	19	14	0,26	[0,23 ;0,30]
l	289	1	0	0,00	[0,23 ;0,30]
i	289	1	0	0,00	[0,23 ;0,30]
n					
F	245	186	107	0,42	[0,38 ;0,46]
é	245	205	95	0,54	[0,49 ;0,57]
m	**245**	**97**	**33**	**0,66**	**[0,61 ;0,70]**
i	245	25	4	0,84	[0,80 ;0,87]
n	245	19	5	0,74	[0,70 ;0,77]
i	245	1	0	0,00	[0,70 ;0,77]
n	245	1	0	0,00	[0,70 ;0,77]

TAB. 3 – *Exemple d'estimation des taux d'influence négatifs et des intervalles de confiance*

3.3 Détermination des zones intéressantes

Au niveau de cette étape, le test d'indépendance entre les variables A et B est réalisé. Cette indépendance peut être évaluée en utilisant la *table de contingence (TC) des variations* . Cette table est calculée en comparant les effectifs observés et les effectifs théoriques, (voir table (4)) ($|effectif\ \ theorique| - |effectif\ \ observe\ |$) Guillaume et Nemmiche (2006); Guillaume (2002). Si ($|effectif\ \ observe| = |effectif\ \ theorique|$) alors A et B sont indépendants, sinon, l'hypothèse d'indépendance est rejetée si cet écart est sigificatif.
La TC des variations de l'association ($sexe$,$salaire$) est présentée table 5. Les cellules nulles dans TC des variations sont ignorées. Les résultats de cette table sont ensuite juxtaposés aux tables des taux d'influence positifs et négatifs pour constituer la source de nos RI. Cette opération est réalisée dans l'étape de coordination des résultats.

3.4 Coordination des résultats

L'étape de coordination des résultats consiste à superposer la TC des variations, (voir exemple dans la table (5) et les tables des taux d'influence positifs et négatifs (voir exemple dans la table (3). Les zones qui coïncident sont alors extraites. En d'autres termes, dans le cas négatif, les cellules retenues sont celles avec un taux d'influence $\geq S^{-}$ et une valeur négative dans le TC des variations. Par symétrie, dans le cas positif, les cellules retenues sont celles avec un taux d'influence $\geq S^{+}$ et une valeur positive dans le TC des variations.

	$Observé$			$Théorique$	
salaire	masculin	féminin	$\sum$	masculin	féminin
[1,00; 6,00]	79	107	186	101	85
[6,10; 11,02]	110	95	205	111	94
[11,11; 16,00]	64	33	97	52	45
[16,14; 20,55]	21	4	25	14	11
[21,25; 26,00]	14	5	19	10	9
[26,29; 26,29]	1	0	1	0	0
[44,50; 44,50]	0	1	1	0	0
$\sum$	289	245	534		

TAB. 4 – *Tableaux de contingence observé et théorique de l'association (sexe,salaire).*

salaire	masculin	féminin
[1,00; 6,00]	-22	22
[6,10; 11,02]	-1	1
[11,11; 16,00]	12	-12
[16,14; 20,55]	7	-7
[21,25; 26,00]	4	-4
[26,29; 26,29]	0	0
[44,50; 44,50]	-1	1

TAB. 5 – *Tableau de contingence des variations de l'association (sexe, salaire).*

Les cellules non valides sont remplacées par 0 dans la table finale, (voir exemple donné dans la table (6) où $S^- = 0.70$ et $S^+ = 0.50$).
À la fin de cette étape, des zones rectangulaires sont formées en regroupant, de façon tabulaire, les cellules voisines présentant le même comportement. Ces zones sont, ensuite utilisées afin de générer les RI.

3.5 Extraction des règles d'influence

En se basant sur les résultats des deux étapes précédentes, les RI positives et négatives sont extraites à partir de la table de coordination. Les règles sont de la forme $A \rightarrow B$ ou $A \rightarrow \neg B$ avec A l'*influent* et B l'*influé*. Les RI générées pour l'association $(sexe,salaire)$ sont listées dans la table (7).

$GenInf$ représente l'intervalle de confiance généralisé pour les taux d'influence $[A; B]$ de la règle. $GenInf$ est construit à partir des intervalles de confiance $[A_i; B_i]$ spécifiques aux différentes valeurs des variables avec i la $i$ème classe de l'association, tel que $A = \min\{A_i\}$ et $B = \max\{B_i\}$.
les RI peuvent alors être présentées sous une forme générale comme dans l'exemple présenté ci-après.
$sexe = masculin \rightarrow (\neg salaire \in [1{,}00; 6{,}00]) \wedge (salaire \in [11{,}11; 16{,}00] \vee salaire \in [16{,}14; 20{,}55] \vee salaire \in [21{,}25; 26{,}00])$ avec $GenInf^- \in [0{,}53; 0{,}61]$ et $GenInf^+ \in [0{,}61; 0{,}87]$.

	$Négatif$		$Positif$	
salaire	masculin	féminin	masculin	féminin
[1,00; 6,00]	0	0	0	0,58
[6,10; 11,02]	0	0	0	0
[11,11; 16,00]	0	0,66	0,66	0
[16,14; 20,55]	0	0,84	0,84	0
[21,25; 26,00]	0	0,74	0,74	0
[26,29; 26,29]	0	0	0	0
[44,50; 44,50]	0	0	0	0

TAB. 6 – *Les taux d'influence valides de l'association (sexe, salaire).*

	sexe/salaire		$GenInf$
$Règles^-$	$sexe = féminin \rightarrow$	$\neg salaire \in [16{,}14; 20{,}55]$ $\vee \neg salaire \in [11{,}11; 16{,}00]$ $\vee \neg salaire \in [21{,}25; 26{,}00]$	[0,61 ;0,87]
	$sexe = masculin \rightarrow$	$\neg salaire \in [1{,}00; 6{,}00]$	[0,53 ; 0,61]
$Règles^+$	$sexe = masculin \rightarrow$	$salaire \in [11{,}11; 16{,}00]$ $\vee salaire \in [16{,}14; 20{,}55]$ $\vee salaire \in [21{,}25; 26{,}00]$	[0,61 ;0,87]
	$sexe = féminin \rightarrow$	$salaire \in [1{,}00; 6{,}00]$	[0,53 ;0,61]

TAB. 7 – *Règles d'influence extraites à partir de l'association (sexe,salaire)*

$sexe = féminin \rightarrow (\neg salaire \in [16{,}14; 20{,}55] \vee \neg salaire \in [11{,}11; 16{,}00] \vee \neg salaire \in [21{,}25; 26{,}00]) \wedge (salaire \in [1{,}00; 6{,}00])$ avec $GenInf^- \in [0{,}61; 0{,}87]$ et $GenInf^+ \in [0{,}53; 0{,}61]$.
Ces règles expriment que le fait d'être un homme augmenterait les chances d'obtenir un salaire entre $11{,}11K$ et $26{,}00K$ et de ce fait, les hommes ont peu de 'chance' d'avoir un salaire entre $1{,}00K$ et $6{,}00K$.
Inverssement, être une femme réduit les chances d'obtenir un salaire entre $11{,}11K$ et $26{,}00K$, ainsi une femme à 'plus de chance' d'avoir un salaire entre $1{,}00K$ et $6{,}00K$.

4 Conclusions et perspectives

Dans ce papier, nous proposons une approche pour l'extraction de règles d'influence quantitatives positives et négatives. La stratégie que nous adoptons combine une méthode pour l'identification des zones d'intérêt grâce aux tableaux de contingence des variations et une mesure d'élagage l'$Influence$ qui analyse le comportement des variables et détermine la nature de leur dépendance. Ce genre d'analyse est très utile, notamment dans notre cadre de travail qui s'appuie sur une sémantique spécifique du type $Influent \rightarrow Influé$. Le système de règles obtenu peut servir de base à l'utilisateur afin d'établir des prédictions et éventuellement lancer

des opérations de préventions.
Nous nous intéressons, à présent, à des mesures statistiques existantes qui, après adaptation à notre sémantique, pourraient lever le voile sur des points critiques de notre méthode -liés notamment aux limitations des seuils minimums- et constituer des éléments de comparaison objectifs avec notre méthode.

Références

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

Antonie, M. et O. Zaane (2004). Mining positive and negative association rules: An approach for confined rules. *Technical Report TR04-07, Dept. of Computing Science, University of Alberta*.

Aumann, Y. et Y. Lindell (1999). A statistical theory for quantitative association rules. *Knowledge Discovery and Data Mining*, 261–270.

Brin, S., R. Motwani, et C. Silverstein (1997). Beyond market baskets: Generalizing association rules to correlations. *ACMSIGMOD Int. Conf. on Management of Data, Tucson, Arizona, USA*, 265–276.

Fukuda, T., Y. Morimoto, S. Morishita, et T. Tokuyama (1996). Data mining using two-dimensional optimized association rules: Scheme, algorithms and visualization. *Proc. of the Int'l Conf. ACMSIGMOD*, 12–23.

Guillaume, S. (2002). Discovery of ordinal association rules. *PAKDD*, 322–327.

Guillaume, S. et L. Nemmiche (2006). Visualisation of attractive and repulsive zones between variables. *The Australasian Data Mining Conference (AusDM), Sydney*.

Hong, T. P., C. S. Kuo, et S. C. Chi (1999). Mining association rules from quantitative data. *Intelligent Data Analysis, Vol. 3, No. 5*, 363–376.

IBM (1996). Ibm intelligent miner user's guide, version 1.1. *IBM, San Jose, CA*.

Lent, B., A. N. Swami, et J. Widom (1997). Clustering association rules. *ICDE*, 220–231.

Macfie, P. B. et M. P. Nufrio (2005). *Applied Statistics for Public Policy*.

Mata, J., J. L. Alvarez, et J. C. Riquelme (2002). An evolutionary algorithm to discover numeric association rules. *ACM symp. Appl. computing SAC'02*, 590–594.

Miller, R. J. et Y. Yang (1997). Association rules over interval data. *ACM SIGMOD Int. Conf. Management of Data*, 452–461.

Murphy, P. et D. Aha (1995). Uci repository ofmachine learning databases. *Machine-readable collection, Dept of Information and Computer Science, Irvine*.

Nemmiche, L. et S. Guillaume (2006). Variables interaction for mining negative and positive quantitative association rules. *The 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), Washington D.C.*.

Rastogi, R. et K. Shim (1999). Mining optimized support rules for numeric attributes. *15th Int. Conf. on Data Engineering,IEEE Computer Society, Sydney, Austrialia*, 206–215.

Savasere, A., E. Omiecinski, et S. B. Navathe (1998). Mining for strong negative associations in a large database of customer transactions. *ICDE*, 494–502.

Srikant, R. et R. Agrawal (1996). Mining quantitative association rules in large relational tables. *ACM SIGMOD Int. Conf. on Management of Data*, 1–12.

Teng, W. G., M. J. Hsieh, et M. S. Chen (2002). On the mining of substitution rules for statistically dependent items. *IEEE Int. Conf. on Data Mining (ICDM'02)*, 442.

Webb, G. I. (2001). Discovering associations with numeric variables. *Knowledge Discovery and Data Mining*, 383–388.

Wu, X., C. Zhang, et S. Zhang (2004). Efficient mining of both positive and negative association rules. *ACM Press, NY, USA*, 381–405.

Yan, P., G. Chen, C. Cornelis, M. D. Cock, et E. E. Kerre (2004). Mining positive and negative fuzzy association rules. *KES 2004*, 270–276.

Yuan, X., B. P. Buckles, Z. Yuan, et J. Zhang (2002). Mining negative association rules. *Computers and Communications (ISCC'02)*, 623–628.

Summary

This article presents an efficient method for mining both positive and negative quantitative influence rules. These influence rules introduce a new semantics which aims to facilitate the analysis of huge data. They allow the visualization of existing interactions between various associations of variables and makes the interpretation of rules easier.
Our approach is based on a strategy combining a tabular regrouping based on contingency tables and a pruning method using an interestingness measure called *Influence*. Thanks to the first step, potentially interesting zones are built and classified. The second step injects selected semantics and evaluates the influence's degree which is produced when introducing a new variable to a set of variables. This step refines the results of the first one, and allows focusing on valid zones according to the specified constraints. Finally, interesting influence rules are generated based on results' juxtaposition of the two key steps of our approach.