Règle et R-règle d'exception en Analyse Statistique Implicative

Régis Gras *, Einoshin Suzuki **, Pascale Kuntz *

*Laboratoire d'Informatique de Nantes Atlantique FRE CNRS 2729
Equipe COD - Connaissances & Décision
Site Ecole Polytechnique de l'Université de Nantes
La Chantrerie BP 60601 44306 Nantes cedex
** Department of Informatics, ISEE, Kyushu University, Japan
regisgra@club-internet.fr , pascale.kuntz@polytech.univ-nantes.fr, suzuki@i.kyushu-u.ac.jp

Résumé. En fouille de règles, certaines situations exceptionnelles défient le bon sens. C'est le cas de la règle R : $a \rightarrow c$ et $b \rightarrow c$ et (a et b) $\rightarrow$ non c. Une telle règle, que nous étudions dans l'article, est appelée règle d'exception. A la suite des travaux précurseurs de E. Suzuki et Y. Kodratoff (1999), qui ont étudié un autre type de règle d'exception, nous cherchons ici à caractériser les conditions d'apparition de la règle R dans le cadre de l'Analyse Statistique Implicative. Nous étendons cette notion aux *R*-règles.

1 Introduction

Depuis les travaux de Agrawal *et al.*, (1993) les règles d'association ont été un modèle très utilisé pour extraire des tendances implicatives dans des bases de données. Rappelons que lorsqu'on dispose d'un ensemble E d'individus décrits par p variables $\{a, b, \ldots.\}$, qui peuvent être des conjonctions de variables atomiques et que l'on supposera ici binaires, une règle d'association $a \rightarrow b$ signifie que si a est vérifiée alors *généralement* b l'est également. Lorsque l'on extrait un ensemble de telles règles *partielles* d'association, il est pertinent de s'interroger sur les « relations » que ces règles entretiennent entre elles. Cette question a été abordée dans la littérature selon différents points de vue. Dans une optique de structuration de l'ensemble des règles, différentes méthodes de classification ont été proposées (e.g. Lent et al., 1997 ; Gras et Kuntz, 2005). Des représentations visuelles bien adaptées permettent également de mettre en évidence des dépendances entre les règles (e.g. Lehn, 2000 ou Couturier et Gras, 2005).

Si l'on étudie localement avec attention ces relations, on peut découvrir une situation qui défie l'intuition. Supposons que l'on ait, entre trois variables (par exemple, des attributs) a, b et c, conjonction de variables binaires dans l'étude présente et vérifiant $a \rightarrow c$ et $b \rightarrow c$. Dans des cas exceptionnels, on n'a pas (a et b) $\rightarrow c$, comme le bon sens nous le suggère, mais (a et b) $\rightarrow \bar{c}$ (où $\bar{c}$ est écrit à la place de non c) Cette dernière règle sera appelée ici *règle d'exception*.

Remarquons que des travaux antérieurs (Suzuki et Kodratoff, 1999 ; Suzuki et Zytkow, 2005) considèrent comme situation d'exception la situation suivante :

$a \rightarrow c$ (dite règle de sens commun), non ($b \rightarrow c'$) (dite règle de référence) et (a *et* b) $\rightarrow c'$ (dite règle d'exception) où $c \neq c'$ et où a et b *sont respectivement des conjonctions(* $a = a_1$ *et* $a_2 \ldots$*et* a_m*)* et *(*$b = b_1$ *et* $b_2 \ldots$ *et* b_p*)*. Notre définition de règle d'exception se distingue ainsi de celle-ci, mais présente comme chez E.Suzuki et Y.Kodratoff, un caractère surprenant.

Or, il existe, comme nous le verrons en donnant des exemples, des situations naturelles où un caractère exceptionnel associe les trois variables. Pour le prendre en compte et en étudier un modèle, nous étendons ici le sens précédent en accentuant ainsi le caractère surprenant (*d'exception*) d'une règle dérivée de deux règles simples.

Pour illustrer ce type de règle, nous faisons référence tout d'abord au cas de l'incompatibilité de groupes sanguins en ce qui concerne le facteur Rhésus.

Certaines femmes, non primo-parturientes, dont les globules rouges sont porteurs de deux allèles Rh- et dont l'immunisation anti-Rh+ est active, possèdent alors le phénotype Rh- (caractère a). Quel que soit le père en général, l'enfant qu'elles portent ne présentera pas, à la naissance, de problème sur le plan sanguin (caractère c). Nous sommes en présence de la règle : $a \rightarrow c$.

Un homme, de génotype Rh+ et Rh+, possède le phénotype Rh+ (caractère b). Quelle que soit la mère en général, l'enfant qu'il engendrera n'aura pas de problème à sa naissance (caractère c). C'est la situation où la règle $b \rightarrow c$ est valide.

En revanche, un couple où la femme est Rh- et remplit les conditions *a* et l'homme est Rh+ (caractère *b*) pourra donner naissance à un enfant qui présentera un risque important du fait de l'incompatibilité Rhésus (caractère $\bar{c}$). Dans des cas exceptionnels, en effet, la mère s'immunisant contre le facteur Rh du fœtus, fabrique des anticorps, qui détruisent les globules rouges de l'enfant. Même si la conjugaison des caractères *a* et *b* est rare, on rencontre cependant la réalisation de la règle, que nous avons appelée « règle d'exception », (*a* et *b*) $\rightarrow \bar{c}$. On sait d'ailleurs que des précautions sont prises pour éviter ce problème dès que sont connus les phénotypes des parents à la faveur d'une prévention adaptée (par exemple l'exsanguino-transfusion).

On trouve une situation comparable d'apparition de règle d'exception dans l'étude des phénomènes d'interférences lumineuses, par exemple dans l'expérience classique des franges de Young (Bruhat G., 1959). La même source lumineuse franchissant deux fentes identiques (*a* et *b*) conduit à des franges d'interférences où alternent des zones d'intensité lumineuse (*c*) variable susceptible de faiblir et/ou s'annuler ($\bar{c}$).

Dans cette communication, nous cherchons à caractériser les situations qui présentent ce type de règles. Nous nous plaçons ici dans le cadre de l'Analyse Statistique Implicative (A.S.I.) qui a montré toute sa pertinence en E.C.D. (Extraction de Connaissances dans les Données) pour la fouille de règles[1].

Dans la première partie, nous rappelons brièvement les principes de base de l'A.S.I. qui nous servent dans la suite. Puis, nous proposons une description ensembliste d'une règle d'exception, suivie d'un exemple numérique dont l'analyse des résultats nous permet d'établir des conjectures sur les conditions favorables à l'apparition d'une règle d'exception. Nous finissons par une caractérisation formelle de ces conditions selon les deux modélisations de tirages les plus utilisées en A.S.I. : le modèle de Poisson et le modèle binomial. Elle s'exprimera en termes de relations algébriques entre les paramètres de la situation.

2 Principes de l'A.S.I. – Rappels

Notons *A* et *B* les sous-ensembles respectifs de *E* d'individus qui vérifient respectivement les variables *a* et *b*. Pour une règle quelconque $a \rightarrow b$, observée dans E, l'A.S.I. consiste à comparer le nombre de contre-exemples $n_{a \wedge \bar{b}}$ à cette règle observés dans $A \cap \bar{B}$, avec le nombre de contre-exemples qui apparaîtraient lors d'un choix aléatoire et indépendant de deux parties *X* et *Y* de *E* de mêmes cardinaux respectifs que *A* et *B* (FIG 1) (cf. (Gras, 2005) et (Lebart et al., 2006)). La variable aléatoire associée est notée $N_{a \wedge \bar{b}}$.

La qualité de la règle $a \rightarrow b$ sera intuitivement d'autant meilleure que $\text{Prob}[N_{a \wedge \bar{b}} > n_{a \wedge \bar{b}}]$ sera proche de 1 : autrement dit, en général, on observe plus de contre-exemples dans des circonstances aléatoires que l'on en a observés dans la contingence. Dans ce cas, le seul hasard conduit donc, en moyenne, à plus de contre-exemples que ce qui est observé.

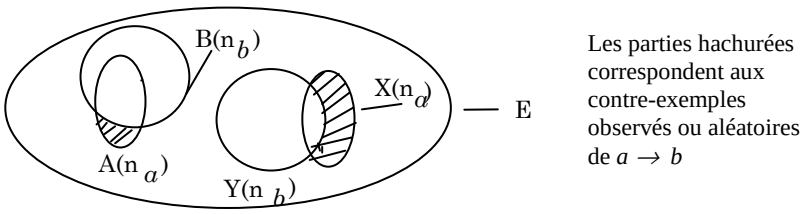


FIG 1 – Représentation ensembliste

La méthode de tirage au hasard de X et Y, dans une hypothèse a priori d'indépendance de *a* et *b*, conduit à différentes options pour la loi de la variable aléatoire $N_{a \wedge \bar{b}}$. Deux modélisations de cette variable sont généralement retenues en A.S.I. conduisant à un modèle de Poisson et à un modèle binomial (Gras, 1979 ; Lerman, 1981a et Lerman et al. 1981b). On centre et on réduit cette variable en la variable $Q(a, \bar{b})$; l'observation contingente, sa réalisation, est $q(a, \bar{b})$. Par exemple, dans le cas de Poisson, on obtient :

[1] cf. Actes de la 3e Rencontre ASI 3 de Palerme, 6-8/10/05, GRIMM, Université de Palerme

$$q(a \wedge b, \bar{c}) = \frac{n_{a \wedge b \wedge \bar{c}} - \frac{n_{a \wedge b} . n_{\bar{c}}}{n}}{\sqrt{\frac{n_{a \wedge b} . n_{\bar{c}}}{n}}}$$

L'intensité d'implication est alors définie par $\varphi(a,b)$ = Prob[$Q(a, \bar{b}) > q(a, \bar{b})$], dont la valeur gaussienne asymptotique, centrée et réduite est :

$$\varphi(a,b) = \frac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{+\infty} e^{-\frac{t^2}{2}} dt \qquad \textbf{(1)}$$

Plus $q(a, \bar{b})$ est négatif, meilleure est la qualité de la règle $a \rightarrow b$.

3 Interprétation et illustration des règles d'exception

Soient *A*, *B*, *C* et $A \cap B$ respectivement les sous-ensembles d'individus de *E* qui satisfont les variables *a*, *b*, *c* et (*a* et *b*). Dans la situation illustrée ici, elles sont binaires, mais l'A.S.I. permet de considérer également d'autres types de variables (Gras 2005).

3.1 Deux approches pour la caractérisation des règles d'exception

Supposons la situation prototypique des règles d'exception : $a \rightarrow c$, $b \rightarrow c$ et (*a* et *b*) $\rightarrow \bar{c}$ (alors que (*a* et *b*) $\rightarrow c$ est de piètre qualité). Elle s'exprime, en termes ensemblistes, par une quasi-inclusion des ensembles d'instances à savoir : *A* et *B* sont presque contenus dans C, mais $A \cap B$ est plutôt contenu dans le complémentaire de *C*. L'illustration ci-dessous rend compte de la situation ensembliste.

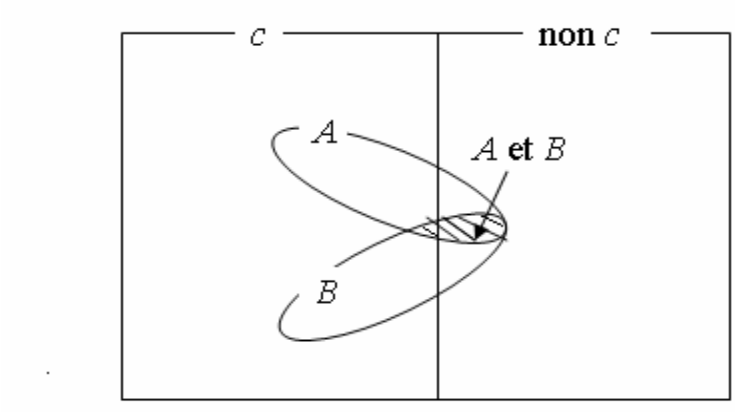


FIG 2 – Apparition d'une règle d'exception ensembliste

Dans le cadre de l'A.S.I., deux approches pourraient nous permettre de mettre en évidence cette situation.

La première approche est basée sur l'analyse de l'intensité d'implication $\varphi(a,b)$ selon la théorie présentée dans le §2. Elle nous permet de conclure au rejet de (*a* et *b*) $\rightarrow c$ et, a contrario, à l'apparition d'une intensité, non négligeable quelquefois, de (*a* et *b*) $\rightarrow \bar{c}$ qui en justifie la prise en compte en tant que règle d'exception. Une représentation en graphe des relations implicatives entre règles élémentaires ci-dessus sera illustrée dans le § 3.2.

La deuxième approche est basée sur l'extension, que nous avons proposée, des règles en *R*-règles (règles de règles) de type $R \rightarrow R'$, où *R* et *R'* sont elles-mêmes des règles (Gras et Kuntz, 2005). Intuitivement, ces règles sont comparables à celles qui apparaissent en mathématiques où un théorème *R* a pour conséquence un autre théorème *R'* ou est suivi d'un corollaire *R'*. Bien évidemment, il ne s'agit que d'une métaphore puisque en A.S.I.

on considère des règles partielles, qui ne sont donc pas strictes et ne relèvent de la logique formelle qu'exceptionnellement. Elles sont construites selon un algorithme récursif utilisant un indice appelé « cohésion ». Celui-ci rend compte de la qualité des liaisons implicatives des variables de la règle *R* avec les variables de la règle *R'*.

Rappelons, qu'en logique formelle, la règle généralisée, règle de règle, ou *R*-règle, $a \Rightarrow (b \Rightarrow c)$, composée de la règle $R_1 = (b \Rightarrow c)$ et $R_2 = (a \Rightarrow R_1)$, est vraie en même temps que (*a* et *b*) $\Rightarrow c$ - donc lui est logiquement équivalente - où les variables *a*, *b* et *c* peuvent être des règles elles-mêmes (Gras et Kuntz., 2005). Or, nous avons vu, dans l'examen des règles élémentaires de la forme $\alpha \rightarrow \beta$, que (*a* et *b*)$\rightarrow \bar{c}$, règle d'exception, est, généralement, en contradiction sémantique avec ($a \rightarrow c$ et $b \rightarrow c$) et que cette conjonction est plutôt formellement compatible avec (*a* et *b*)$\rightarrow c$**.**

De la même façon, la R-règle $a \rightarrow (b \rightarrow c)$ est en contradiction formelle avec (*a* et *b*) $\rightarrow \bar{c}$. Mais comme nous sommes dans le cadre de l'A.S.I. où les règles sont partielles, cette dernière règle peut apparaître bien qu'elle soit inattendue. Nous dirons alors, comme précédemment, que (*a* et *b*) $\rightarrow \bar{c}$ est **une règle d'exception** de la R-règle $a \rightarrow (b \rightarrow c)$. Un arbre hiérarchique illustre dans le paragraphe suivant cette approche par des R-règles.

3.2 Exemple numérique

Nous avons construit un fichier fictif de 200 sujets (cf. un tableau partiel en Annexe où nous en montrons la construction), sujets sur lesquels nous observons les variables binaires : *a*, *b* , $a \wedge b$, *c* et $\bar{c}$. Les valeurs associées des différentes intensités sont données dans TAB 1. Elles sont obtenues par le logiciel CHIC (Couturier et Gras, 2005) qui permet les calculs et les représentations graphiques des ensembles de règles extraites des instances,

	a	*b*	*c*	$\bar{c}$	$a \wedge b$
a	0	.79	.89	.08	.89
b	.79	0	.84	.10	.88
c	.68	.67	0	0	.36
$\bar{c}$	.32	.33	0	0	.64
$a \wedge b$	1.00	1.00	.03	.97	0

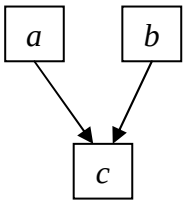
TAB 1 –.Intensités d'implication associées à un jeu de données

Notons les fréquences des occurrences des variables : $n_a = n_b = 12$; $n_{a \wedge b} = 7$; $n_c = 50$. Les intensités d'implication associées sont :

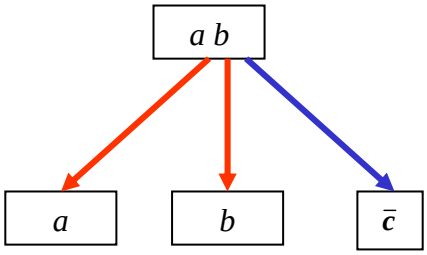
$\varphi(a,c) = 0.89$; $\varphi(b,c) = .84$, $\varphi((a \text{ et } b), c) = 0.03$

alors que $\varphi((a \text{ et } b), \bar{c}) = 0.97$; ce qui confirme la présence d'une règle d'exception.

a) Selon la première approche de règles élémentaires, une analyse par CHIC sur le tableau complet donne le graphe implicatif (Figure 4.a) et l'on constate la bonne qualité d'implication de *a* et de *b* sur *c*. On constate aussi que l'on a bien $a \rightarrow c$ et $b \rightarrow c$. Lorsque CHIC conjoint les variables, on obtient cette fois le phénomène lié à l'existence d'une règle d'exception (Figure 4.b).



4.a- Règles simples



4.b- Apparition de la règle d'exception

FIG 4– Graphe implicatif avec les règles simples puis conjointes

.

D'une façon générale en A.S.I., trois conditions nous semblent favorables à l'apparition d'une règle d'exception de $a \wedge b$ sur non c :

Une certaine qualité d'implication de a et de b sur c ; cette condition de bon sens conduit à ce que la règle $(a \wedge b) \rightarrow c$ soit attendue et non pas $(a \wedge b) \rightarrow \bar{c}$ qui en définitive va l'être;

Une mauvaise qualité de ressemblance entre $(a \wedge b)$ et c ($n_{a \wedge b \wedge c}$ est faible) ;

Une bonne qualité de ressemblance entre a et b si le référentiel devient $\bar{c}$ ($n_{a \wedge b \wedge \bar{c}}$ est grand relativement à $n_{a \wedge b}$).

b) Selon la deuxième approche, relativement à cet exemple numérique, $a \rightarrow b$, faiblement (.81), et $a \rightarrow c$ et $b \rightarrow c$ un peu plus fortement (.85 et .89). Par suite, la R-règle $a \rightarrow (b \rightarrow c)$ est validée par l'ASI et restituée au moyen du logiciel CHIC. Les arbres suivants illustrent respectivement d'une part la règle généralisée $a \rightarrow (b \rightarrow c)$ où l'on voit que a et b n'ont aucune relation implicative avec non c, d'autre part que la règle d'exception (a et b) $\rightarrow \bar{c}$ est obtenue lorsque l'on conjoint a et b.

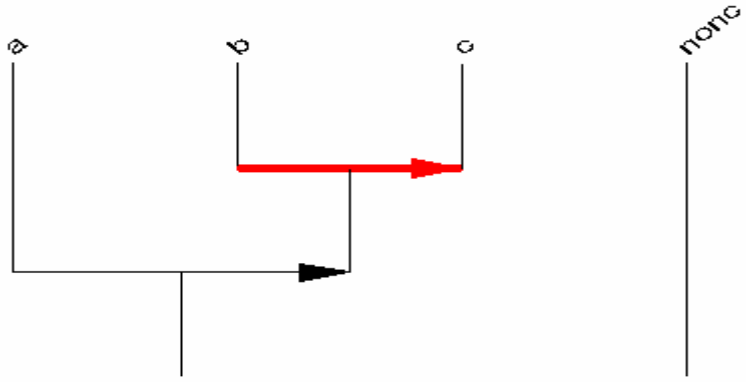


FIG 5-Représentation hiérarchique de la R-règle $a \rightarrow (b \rightarrow c)$

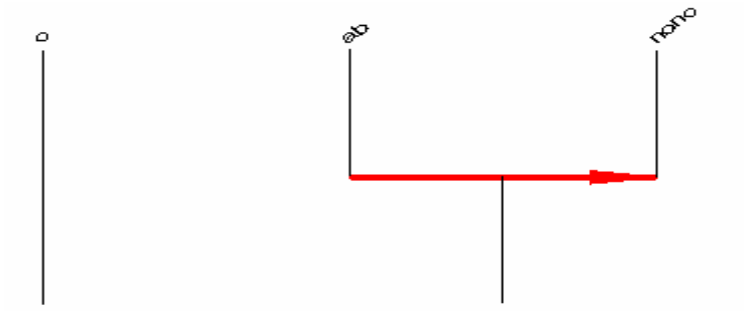


FIG 6- Représentation hiérarchique de la R-règle d'exception (a et b)$\rightarrow \bar{c}$

Autrement dit, l'absence de cohérence entre l'arbre extrait des données (FIG.5) et l'arbre après conjonction des variables (FIG.6) témoigne de l'apparition de la règle d'exception. Celle-ci a pu être observée de façon analogue à travers l'absence de cohérence entre les deux représentions des graphes implicatifs.

Ainsi, comme pour l'approche graphique ci-dessus, la double construction de la hiérarchie implicative : variables élémentaires (FIG 4, a) puis variables conjointes (FIG 4, b), permet par examen de la non-cohérence de repérer l'existence d'une R-règle d'exception.

4 Relation entre les intensités d'implication de a $\wedge$ b sur c et sur non c

Rappelons que, en A.S.I., nous modélisons l'implication de a sur b de deux manières : par une loi de Poisson de paramètre $\lambda = n_a n_{\bar{b}} / n$;

par une loi binomiale de paramètres n et $p = n_a n_{\bar{b}} / n.n$ . .

Une modélisation hypergéométrique est écartée car elle n'induit pas de différence entre une implication et sa réciproque (Gras et al., 1996b).

Etablissons pour chacun de ces deux modèles retenus les intensités d'implication de la conjonction $a \wedge b$ sur les variables c et non c (encore notée $\bar{c}$). Nous utiliserons la relation simple : $n_{a\wedge b\wedge \bar{c}} = n_{a\wedge b} - n_{a\wedge b\wedge c}$.

4.1 Modèle de Poisson

a) Première approche par règles élémentaires, dans ce modèle, pour respectivement l'implication $a \wedge b \rightarrow \bar{c}$ et l'implication $a \wedge b \rightarrow c$, les indices $q_1(a \wedge b, c)$ et $q_2(a \wedge b, \bar{c})$ sont

$$q_1 = \frac{n_{a\wedge b\wedge c} - \frac{n_{a\wedge b}.n_c}{n}}{\sqrt{\frac{n_{a\wedge b}.n_c}{n}}} \qquad \text{et} \quad q_2 = \frac{n_{a\wedge b\wedge \bar{c}} - \frac{n_{a\wedge b}.n_{\bar{c}}}{n}}{\sqrt{\frac{n_{a\wedge b}.n_{\bar{c}}}{n}}} \qquad \textbf{(2)}$$

Pour que l'implication $a \wedge b \rightarrow \bar{c}$ soit de bonne qualité, il est nécessaire que q_1 soit négatif.

En effet, le nombre de contre-exemples observé $n_{a\wedge b\wedge c}$ doit être inférieur à celui auquel seul le hasard pourrait conduire, dans l'hypothèse d'indépendance de $a \wedge b$ et de c, soit la moyenne $\frac{n_{a\wedge b}.n_c}{n}$.

On retrouve ici un argument justifiant le cas (2) du paragraphe précédent. Des définitions (2) on déduit :

$$q_1 = \frac{n_{a\wedge b} - n_{a\wedge b\wedge \bar{c}} - \frac{n_{a\wedge b}.n_c}{n}}{\sqrt{\frac{n_{a\wedge b}.n_c}{n}}} = - \frac{n_{a\wedge b\wedge \bar{c}}.n - n_{a\wedge b}(n - n_c)}{\sqrt{n.n_{a\wedge b}.n_c}}$$

$$= \frac{n_{a\wedge b\wedge \bar{c}} - \frac{n_{a\wedge b}.n_{\hat{c}}}{n}}{\sqrt{\frac{n_{a\wedge b}.n_c}{n}}} = -q_2 . \frac{\sqrt{\frac{n_{a\wedge b}.n_{\bar{c}}}{n}}}{\sqrt{\frac{n_{a\wedge b}.n_c}{n}}}$$

Finalement : $q_1 = - q_2 \sqrt{\frac{n_{\bar{c}}}{n_c}}$ soit $\frac{q_1}{q_2} = - \sqrt{\frac{n_{\bar{c}}}{n_c}}$.

q_1 et q_2 sont bien de signes opposés, ce qui est conforme à l'intuition. Mais de plus, l'amplitude de la positivité de q_2 induit celle de la négativité de q_1.

Au sens de l'intensité d'implication (classique), pour que la règle $a \wedge b \rightarrow \bar{c}$ soit considérée comme une exception et apparaisse, la différence $\varphi(a \wedge b, \bar{c}) - \varphi(a \wedge b, c)$ suivante doit être positive et suffisamment grande:

$$\frac{1}{\sqrt{2\pi}} \int_{q_1}^{+\infty} e^{-\frac{t^2}{2}} dt - \frac{1}{\sqrt{2\pi}} \int_{-q_1\sqrt{\frac{n_c}{n_{\bar{c}}}}}^{+\infty} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_{q_1}^{-q_1\sqrt{\frac{n_c}{n_{\bar{c}}}}} e^{-\frac{t^2}{2}} dt \qquad \textbf{(3)}$$

Conséquence 1. Il y a donc apparition de règles d'exception :

- lorsque q_1 est négatif, c'est-à-dire lorsque $\frac{n_{a\wedge b\wedge c}}{n} < \frac{n_{a\wedge b}}{n}.\frac{n_c}{n}$ (sous-indépendance) ;

- de qualité d'autant meilleure que l'ensemble C des instances satisfaisant *c* est supérieur à celui qui satisfont sa négation non *c*. La force de l'intensité d'exception sera à la mesure de la valeur de l'intégrale gaussienne sur l'intervalle $[q_1 ; -q_1\sqrt{\frac{n_c}{n_{\bar{c}}}}]$. De même, règle attendue et règle d'exception coïncident lorsque $q_1 = q_2 = 0$.

Ainsi, c'est à l'occasion de l'indépendance de $a \wedge b$ et $\bar{c}$ et donc de $a \wedge b$ et c que disparaît la règle d'exception.

b) Dans le cadre de l'approche par R-règles, considérant la règle $(b \rightarrow c)$ comme une variable binaire, c'est-à-dire prenant respectivement la valeur 0 lorsque *b*=1 alors que *c*=0 et la valeur 1 dans les autres cas, les contre-exemples à la règle sont en nombre $n_{b\wedge\bar{c}}$. Dans ces conditions, les contre-exemples à la R-règle $a \rightarrow (b \rightarrow c)$ apparaissent lorsque *a* prend la valeur 1 alors que $(b \rightarrow c)$ prend la valeur 0, c'est-à-dire lorsque $b \wedge \bar{c}$ prend la valeur 1. Par suite, le nombre de ces contre-exemples est $n_{a\wedge b\wedge\bar{c}}$ et l'indice d'implication associé à la R-règle, dans une **modélisation de Poisson** de l'implication, est alors :

$$q_3 = \frac{n_{a\wedge b\wedge\bar{c}} - \frac{n_a.n_{b\wedge\bar{c}}}{n}}{\sqrt{\frac{n_a.n_{b\wedge\bar{c}}}{n}}} = \frac{n.n_{a\wedge b\wedge\bar{c}} - n_a.n_{b\wedge\bar{c}}}{\sqrt{n.n_a.n_{b\wedge\bar{c}}}}$$

On constate que cet indice est différent de celui associé à la règle $a \wedge b \rightarrow c$ qui est la règle élémentaire attendue de la conjonction $a \rightarrow c$ et $b \rightarrow c$. En conséquence, les indicateurs qui nous permettront de prévoir l'existence d'une règle d'exception dans cette approche hiérarchique seront différents de ceux qui nous permettent d'anticiper l'exception dans l'approche par le graphe implicatif.

Rappelons alors que l'indice d'implication de $(a \text{ et } b) \rightarrow \bar{c}$ est : $q_1 = \frac{n_{a\wedge b\wedge c} - \frac{n_{a\wedge b}.n_c}{n}}{\sqrt{\frac{n_{a\wedge b}.n_c}{n}}} =$

$$\frac{n.n_{a\wedge b\wedge c} - n_{a\wedge b}.n_c}{\sqrt{n.n_{a\wedge b}.n_c}}$$

On démontre, par transformation des deux indices, que q_1 est négatif (donc la règle d'exception est valide) alors que q_3 est positif (donc la règle attendue n'apparaît pas) si et seulement si :

$$\frac{n_{a\wedge b}}{n_a} > \frac{n_{b\wedge\bar{c}}}{n_{\bar{c}}}$$

C'est-à-dire si la fréquence conditionnelle de *b* dans *a* est supérieure à sa fréquence dans $\bar{c}$. Inversement, si cette inégalité est observée dans l'autres sens, la règle attendue apparaît alors que la règle d'exception n'existe pas.

Dans l'exemple numérique qui illustre la règle d'exception $a \wedge b \rightarrow \bar{c}$, nous avons :

d'une part : $n_{a\wedge b}$ =7, n_b = 24, soit $\frac{n_{a\wedge b}}{n_a}$ = 0,29, et,

d'autre part : $n_{b\wedge\bar{c}}$ = 8, $n_{\bar{c}}$ = 100 soit $\frac{n_{b\wedge\bar{c}}}{n_{\bar{c}}}$ = 0,08. L'inégalité est bien satisfaite.

Ce résultat analytique est différent de celui obtenu par l'approche graphique ce qui confirme bien la différence de signification des deux représentations de l'implication.

4.2 Modèle binomial

Posons q'$_1$ et q'$_2$ respectivement les indices respectifs d'implication de $a \wedge b \rightarrow \bar{c}$ et $a \wedge b \rightarrow c$, lorsque le modèle de tirage aléatoire des parties *A*, *B* et *C* est binomial. Dans ce cas, par un calcul comparable au modèle précédent, on obtient

$$\frac{q'_1}{q'_2} = - \sqrt{\frac{n_{\bar{c}}\,(n^2 - n_{a\wedge b}.n_{\bar{c}})}{n_c\,(n^2 - n_{a\wedge b}.n_c)}} \qquad \textbf{(4)}$$

Posons

$$k(a,b,c) = \left[\frac{\left(1 - \frac{n_{a\wedge b}.n_{\bar{c}}}{n^2}\right)}{\left(1 - \frac{n_{a\wedge b}.n_c}{n^2}\right)} \right]^{\frac{1}{2}} \qquad \textbf{(5)}$$

Donc,

$$\frac{q'_1}{q'_2} = - \sqrt{\frac{n_{\bar{c}}}{n_c}}.\, k(a,b,c) \qquad \textbf{(6)}$$

et la différence $\varphi(a \wedge b, \bar{c})$ - $\varphi(a \wedge b, c)$ entre les intensités d'implication est

$$\frac{1}{\sqrt{2\pi}} \int_{q'_1}^{+\infty} e^{-\frac{t^2}{2}} dt \; - \frac{1}{\sqrt{2\pi}} \int_{-q'_1\sqrt{\frac{n_c}{n_{\bar{c}}}}.\,k\,(a,b,c)}^{+\infty} e^{-\frac{t^2}{2}} dt \qquad = \frac{1}{\sqrt{2\pi}} \int_{q'_1}^{-q'_1\sqrt{\frac{n_c}{n_{\bar{c}}}}.\,k(a,b,c)} e^{-\frac{t^2}{2}} dt \qquad \textbf{(7)}$$

Conséquence 2. Pour le modèle binomial, la différence entre les intensités d'implication sera non seulement fonction du rapport $\frac{n_c}{n_{\bar{c}}}$ mais aussi de *k*(*a*,*b*,*c*) (**5**). Ce coefficient est d'autant plus grand et renforce ainsi l'effet du rapport $\frac{n_c}{n_{\bar{c}}}$ que $1 - \frac{n_{a\wedge b}.n_{\bar{c}}}{n^2} >> 1 - \frac{n_{a\wedge b}.n_c}{n^2}$, c'est-à-dire $\frac{n_{a\wedge b}}{n}.\frac{n_c}{n} >> \frac{n_{a\wedge b}}{n}.\frac{n_{\bar{c}}}{n}$. Les deux membres de cette inégalité ne sont autres que, de gauche à droite, les probabilités respectives du nombre de contre-exemples aléatoires - dans le modèle binomial où les variables $a \wedge b$ et c seraient indépendantes - des implications $a \wedge b \rightarrow \bar{c}$ et $a \wedge b \rightarrow c$. Ainsi, plus on s'attend à une réfutation de $a \wedge b \rightarrow \bar{c}$, au vu de $n_{a\wedge b}$ et de $\bar{c}$, plus le caractère surprenant, *exceptionnel*, de cette règle est manifeste par le constat de la modicité des contre-exemples observés à savoir $n_{a\wedge b\wedge c}$. Ceux-ci la valident au détriment de $a \wedge b \rightarrow c$ (argument favorable à la conjecture 3 du § 3.2). L'inégalité montre la « contribution active à l'exception » au rapport $\frac{n_c}{n_{\bar{c}}}$, contribution qu'apportent les instances, de cardinal $n_{a\wedge b}$ dans celles de cardinal $n_{\bar{c}}$.

Cette conséquence 2, liée au modèle binomial, nous apparaît donc plus riche que la conséquence 1 car elle nous fournit une relation de contrôle entre les paramètres plus fine du caractère d'exception que dans le modèle de Poisson. Ce phénomène, certes lié au nombre de paramètres de définition du modèle binomial, le gratifie cependant d'un intérêt que le logiciel CHIC permet d'exploiter à travers l'offre de son menu.

Remarque : A titre de comparaison, intéressons-nous à un autre indice de mesure de qualité de règle, la *confiance c*, qui est à la base des principaux autres indices de qualité (Lenca et al. 2004). Elle s'exprime ainsi :

c($a \rightarrow c$) = $\frac{n_{a\wedge c}}{n_a}$ (souvent notée : $\frac{\Pr[a \wedge c]}{\Pr[a]}$, autrement dite probabilité conditionnelle de *c* sachant *a*).

La relation entre les règles que nous avons examinées est alors :

$c(a \wedge b \rightarrow \overline{c}) = \frac{n_{a \wedge b \wedge \overline{c}}}{n_{a \wedge b}} = 1 - \frac{n_{a \wedge b \wedge c}}{n_{a \wedge b}} = 1 - c(a \wedge b \rightarrow c)$

La règle d'exception a pour mesure le complément à 1 de la règle attendue. Ainsi, elle est indépendante des valeurs des occurrences.

5 Conclusion

Lorsque deux variables impliquent une 3ème, que leur conjonction implique plutôt la négation de cette 3ème, nous considérons que cette règle est d'exception, en un sens voisin mais différent de celui de E.Suzuki et Y.Kodratoff (1999). Nous avons étudié et illustré par un exemple numérique et un exemple de génétique, l'expression de ce caractère exceptionnel. Puis nous avons précisé les relations entre les paramètres des variables dans les deux modélisations selon lesquelles est construite l'Analyse Statistique Implicative : un modèle de Poisson et un modèle binomial, l'un et l'autre convergeant vers le même modèle gaussien.

Nous avons évoqué une approche complémentaire pour la détection de ces règles qui se base sur les travaux menés ces dernières années sur les R-règles (Gras et Kuntz, 2005). La construction associée d'une hiérarchie implicative n'a pas été initialement développée dans ce but. Cependant, elle constitue une piste à explorer tant d'un point vue algorithmique que méthodologique concernant l'interprétation de ce qui pourraient être des « R-règles d'exception ».

Références

Agrawal R., Imiliensky T. et Swami A. (1996). Mining association rules between sets of items in large databases. *Proc. of the ACM SIGMOD'93*, p. 679-696, AAAI Press.

Bruhat G. (1959), *Optique*, Masson, Paris.

Couturier R, Gras R. (2005). CHIC : Traitement de données avec l'analyse implicative, *Extraction et Gestion des Connaissances, Volume I1, RNTI, Cepadues, Paris*, p.679-684.

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

Gras, R. et H. Ratsimba-Rajohn. (1996a). Analyse non symétrique de données par l'implication statistique. *RAIRO-Recherche Opérationnelle, 30(3)*, 217-232.

Gras, R., S. Ag Almouloud, M. Bailleul , A. Larher., M. Polo, H. Ratsimba-Rajohn et A. Totohasina, (1996b*). L'implication Statistique, La Pensée Sauvage, Grenoble*.

Gras R., Kuntz P. et Briand H. (2001). Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données, Mathématiques et Sciences Humaines, n° 154-155, pp. 9-29.

Gras R., Kuntz P. et Régnier J.-C. (2004). Significativité des niveaux d'une hiérarchie orientée en analyse statistique implicative, Numéro spécial Classification, Revue des Nouvelles Technologies de l'Information, Cépaduès.

Gras R. (2005). Panorama du développement de l'A.S.I. à travers des situations fondatrices, *Actes de la 3ème Rencontre Internationale A.S.I., Supplément n° 15 de la Revue « Quaderni di Ricerca in Didattica »*, p. 9-33, Université de Palerme**.**

Gras, R. et Kuntz P. (2005). Discovering R-rules with a directed hierarchy, *Soft Computing, A fusion of Foundations, Methodologies and Applications, vol. 10, n°5*, p. 453-460.

Gras R., Kuntz P., Suzuki E. (2007). Une règle d'exception en Analyse Statistique Implicative, *Extraction des Connaissances (EGC'07),Volume1, RNTI-E-9, Cépaduès Editions, ,* p.87-98, ISBN : 1764-1667

Kuntz, P. (2005). Classification hiérarchique orientée en ASI , *Actes de la 3ème Rencontre Internationale A.S.I., Supplément n° 15 de la Revue « Quaderni di Ricerca in Didattica »*, p.53-62, Université de Palerme.

Lehn R.(2000). *Un système interactif de visualisation et de fouille de règles pour l'extraction de connaissances dans une base de données*. Thèse de doctorat, Université de Nantes.

Lebart L., Piron M. et Morineau A.(2006). *Statistique exploratoire multidimensionnelle.* 4ème édition, Science sup, Dunod.

Lenca P., Meyer P., Vaillant P., Picouet P. et Lallich S., (2004). Evaluation et analyse multi-critères de qualité des règles d'association, Mesures de qualité pour la fouille de données, RNTI-E-1,Cépaduès, p. 219-246.

Lent B., Swami A.N. et Widow J. (1997). Clustering association rules. *Proc. of the 13th Int. Conf. on Data Engineering*, p. 220-231.

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

Lerman, I.C., R. Gras et H. Rostam (1981b). Elaboration et évaluation d'un indice d'implication pour données binaires , *Mathématiques et Sc. Humaines, n°74*, 5-35.

Suzuki E. et Kodratoff.Y. (1999). Discovery of surprising exception rules based on intensity of implication. *Principles of data mining and knowledge discovery science. Springer*, p 184-195, 1999.

Suzuki E. et Zytkow J. (2005). Unified algorithm for undirected discovery of exception rules, *Int. J. of Intelligent Systems, vol. 20, Wiley*, p. 673-694.

Summary

In Rule Mining some exceptional situations are contrary to the common sense: $a \rightarrow c$ and $b \rightarrow c$ and (a et b) $\rightarrow$ non c. These rules are called exception rules. Following the precursory work of E. Suzuki and Y. Kodratoff, we study the conditions in which these rules appear in the framework of the Implicative Statistical Analysis (S.I.A.). We extend this notion to *R*-rules.

Annexe

Nous avons construit un fichier fictif de 200 sujets sur lesquels nous observons les variables binaires : a, b, $a \wedge b$, c et non c. Voici les 20 premières lignes du tableau. Ce sont les 4 premières qui vont principalement intervenir dans l'apparition de la règle d'exception **:**

sujets	**a**	**b**	**a $\wedge$ b**	**c**	**nonc**
1	1	1	1	0	1
2	1	1	1	0	1
3	1	1	1	0	1
4	0	1	0	0	1
5	1	0	0	1	0
6	0	1	0	1	0
7	1	0	0	1	0
8	0	1	0	1	0
9	1	0	0	1	0
10	0	1	0	1	0
11	1	0	0	1	0
12	0	1	0	1	0
13	1	0	0	1	0
14	0	1	0	1	0
15	1	0	0	1	0
16	0	1	0	1	0
17	1	0	0	1	0
18	0	1	0	1	0
19	1	0	0	1	0
20	0	1	0	1	0
….	….	….	….	….	….