Une base pour les règles d'association valides au sens de la mesure de qualité $\mathrm{M_{GK}}$

Daniel R. Feno [*,**], Jean Diatta[*], André Totohasina[**]

[*]Université de la Réunion
15 avenue René Cassin - BP 7151
97715 Saint-Denis messag cedex 9-France
fenodaniel2@yahoo.fr, jean.diatta@univ-reunion.fr
[**]Université d'Antsiranana- BP O
201 Antsiranana-Madagascar
totohasina@yahoo.fr

Résumé. Ce papier concerne les règles d'association valides au sens de la mesure de qualité $\mathrm{M_{GK}}$ [1]. D'une part, $\mathrm{M_{GK}}$ est normalisée en ce sens que ses valeurs sont comprises entre -1 et $+1$ et reflètent les situations de référence telles que l'incompatibilité, la dépendance négative, l'indépendance, la dépendance positive et l'implication logique entre la prémisse et le conséquent d'une règle. D'autre part, ses propriétés permettent non seulement de considérer les règles positives et négatives (à droite, à gauche, à gauche et à droite), mais aussi de se restreindre uniquement aux règles positives et à celles négatives à droite. Ainsi, nous proposons une base pour les règles positives exactes, une base pour les règles négatives exactes, une base pour les règles positives approximatives et une base pour les règles négatives approximatives. La réunion de ces quatre bases constitue une base pour toutes les règles d'association (positives et négatives) valides au sens de $\mathrm{M_{GK}}$.

1 Introduction

La fouille de données est un domaine de recherche actif dont l'importance n'a cessé de croître ces dernières années, du fait de son rôle comme outil approprié pour faire face à la croissance explosive de la taille de données stockées. Plusieurs techniques de fouille de données ont été proposées dans la littérature. La fouille des règles d'association en est l'une des plus populaires. Originellement appliquée dans le domaine de Marketing (Agrawal et al., 1993), de nos jours, la fouille des règles d'association s'applique dans différents domaines telles des bases de données médicales (Azé et al., 2003), des bases de données spatiales (Salleb, 2003), des taxonomies (Gras et Totohasina, 1995). Les règles d'association sont utiles pour la découverte de relations au sein de très grandes bases de données. Plusieurs algorithmes de fouille de règles d'association, fondés sur les mesures de qualité support et confiance, ont été proposés dans la littérature : APRIORI (Agrawal et al., 1993), CLOSE (Pasquier et al., 1999), CLOSET

[1] $\mathrm{M_{GK}}$: Mesure de Guillaume-Kenchaff.

(Pei et al., 2000). Toutefois, l'ensemble des règles d'association valides au sens d'une mesure de qualité comporte souvent un très grand nombre de règles dont plusieurs peuvent être redondantes par rapport à des axiomes d'inférence donnés. Ainsi d'un point de vue informatif, il peut s'avérer intéressant de n'en générer qu'une base, c'est-à-dire un ensemble minimal à partir duquel toutes les règles valides peuvent être retrouvées par application de ces axiomes d'inférence.
Notons que la mesure Confiance (Agrawal et al., 1993), souvent utilisée par les différentes méthodes de la fouille des règles d'association, autorisent certaines règles non pertinentes telles que la prémisse et le conséquent sont indépendants (Lallich et Teytaud, 2004). Le présent travail concerne la fouille des règles d'association au sens de la mesure de qualité $\mathrm{M_{GK}}$ introduite indépendamment dans (Guillaume, 2000) et dans (Wu et al., 2004), et dont les propriétés mathématiques ont été étudiées dans (Totohasina, 2003). Contrairement à la Confiance, $\mathrm{M_{GK}}$ ne séléctionne que des règles telles que la prémisse favorise le conséquent. Par ailleurs, $\mathrm{M_{GK}}$ vérifie les trois principes de Piatetsky-Shapiro (1991) tout en reflétant les situations de référence en cas de l'implication totale et l'incompatiblité entre la prémisse et le conséquent, ce qui n'est pas le cas pour la mesure de Piatetsky-Shapiro (1991). Dans le présent article, nous proposons une base pour les règles d'association valides au sens de la mesure $\mathrm{M_{GK}}$. Les propriétés de $\mathrm{M_{GK}}$ permettent non seulement de considérer les règles positives et négatives (à droite, à gauche, à gauche et à droite), *i.e.*, les règles portant sur les motifs et négation des motifs ou les règles portant sur les motifs négatifs, mais aussi de se restreindre uniquement aux règles positives et á celles négatives à droite. Ainsi, nous proposons une base pour les règles positives exactes, une base pour les règles négatives exactes, une base pour les règles positives approximatives et une base pour les règles négatives approximatives. La réunion de ces quatre bases constitue une base pour toutes les règles d'association (positives et négatives) valides au sens de la mesure de qualité $\mathrm{M_{GK}}$.

Le reste du papier est organisé de la façon suivante. Dans la Section 2, nous présentons quelques généralités sur les règles d'association. Section 3 concerne les mesures de qualité des règles d'association ainsi que les propriétés de $\mathrm{M_{GK}}$. Les quatre bases partielles constituant la base pour les règles d'association valides au sens de $\mathrm{M_{GK}}$ sont présentées dans la Section 4. Nous terminons par une courte conclusion dans la Section 5.

2 Règles d'association

Dans cet article, nous nous plaçons dans le cadre d'un contexte binaire $\mathbb{K} = (\mathcal{E}, \mathcal{V})$ où $\mathcal{E}$ est un ensemble fini d'entités et $\mathcal{V}$ un ensemble fini de variables booléennes définies sur $\mathcal{E}$. Les sous ensembles de $\mathcal{V}$ seront appelés *motifs positifs* ou tout simplement *motifs*.

Etant donné un motif X :

- X' désignera l'ensemble des entités vérifiant le motif X, *i.e.*, $X' = \{e \in \mathcal{E} : \forall x \in X, x(e) = 1\}$.
- $\overline{X}$ désignera la négation de X, *i.e.*, $\overline{X}(e) = 1$ si et seulement s'il existe $x \in X$ tel que $x(e) = 0$ ($\overline{X}$' est le complémentaire de X'). Si X est un motif positif, alors $\overline{X}$ sera appelé *motif négatif*.

Le Tableau 1 présente un contexte binaire $\mathbb{K} = (\mathcal{E}, \mathcal{V})$, où $\mathcal{E} = \{e_1, e_2, e_3, e_4, e_5\}$ et $\mathcal{V} = \{A, B, C, D, E\}$. Pour $X = \{B, C\}$, on a $X' = \{e_2, e_3, e_5\}$ et $\overline{X}' = \{e_1, e_4\}$.

Etant donnés deux motifs positifs X et Y de $\mathbb{K}$, nous posons les définitions suivantes :

$\mathcal{E} \backslash \mathcal{V}$	A	B	C	D	E
e_1	1	0	1	1	0
e_2	0	1	1	0	1
e_3	1	1	1	0	1
e_4	0	1	0	0	1
e_5	1	1	1	0	1

TAB. 1 – *Contexte binaire*

- une *règle d'association positive* est un couple (X, Y) de motifs, noté $X \rightarrow Y$;
- une *règle d'association négative à droite* est un couple de la forme $(X, \overline{Y})$, noté $X \rightarrow \overline{Y}$;
- une *règle d'association négative à gauche* est un couple de la forme $(\overline{X}, Y)$, noté $\overline{X} \rightarrow Y$;
- une *règle d'association bilatéralement négative* est un couple de la forme $(\overline{X}, \overline{Y})$, noté $\overline{X} \rightarrow \overline{Y}$.

Pour une règle d'association $X \rightarrow Y$ où X et Y sont des motifs positifs ou négatifs, X est appelé la *prémisse* de la règle et Y son *conséquent*. La validité des règles d'association est évaluée par une (ou plusieurs) mesure(s) de qualité pour ne retenir que les règles d'association pertinentes au sens de cette (ou de ces) mesure(s).

3 Mesure de qualité des règles d'association

Une *mesure de qualité* des règles d'association est une application μ qui associe une valeur réelle à chaque règle d'association. Plusieurs mesures de qualité des règles d'association ont été proposées dans la littérature. Les plus connues d'entre elles sont sans doute le support et la confiance (Agrawal et al., 1993). Pour un ensemble E, désignons par $|E|$ la cardinalité de E. Le support d'un motif X est le nombre réel défini par $\mathrm{Supp}(X) = \frac{|X'|}{|\mathcal{E}|}$. Notons p la probabilité intuitive définie sur $(\mathcal{E}, \mathcal{P}(\mathcal{E}))$ par $p(E) = \frac{|E|}{|\mathcal{E}|}$ pour $E \subseteq \mathcal{E}$. Alors, le support d'un motif X peut s'écrire en fonction de p comme $\mathrm{Supp}(X) = p(X')$. Le support d'une règle $X \rightarrow Y$, noté $\mathrm{Supp}(X \rightarrow Y)$, est défini par :

$$\mathrm{Supp}(X \rightarrow Y) = \mathrm{Supp}(X \cup Y) = p((X \cup Y)') = p(X' \cap Y').$$

Il indique la proportion d'entités vérifiant à la fois la prémisse et le conséquent de la règle. La confiance d'une règle $X \rightarrow Y$, notée $\mathrm{Conf}(X \rightarrow Y)$, est définie par :

$$\mathrm{Conf}(X \rightarrow Y) = \frac{\mathrm{Supp}(X \rightarrow Y)}{\mathrm{Supp}(X)} = \frac{p(X' \cap Y')}{p(X')} = p(Y'|X').$$

où $p(Y'|X')$ est la probabilité conditionnelle de Y' sachant X'. Elle indique la proportion d'entités vérifiant le conséquent parmi celles vérifiant la prémisse.

Dans ce papier, nous nous intéressons à la mesure de qualité $\mathrm{M_{GK}}$, introduite indépendamment dans (Guillaume, 2000) et dans (Wu et al., 2004), et définie par :

$$\mathrm{M_{GK}}(X \rightarrow Y) = \begin{cases} \frac{p(Y'|X') - p(Y')}{1 - p(Y')} & si\ p(Y'|X') \geq p(Y') \\ \frac{p(Y'|X') - p(Y')}{p(Y')} & si\ p(Y'|X') \leq p(Y'). \end{cases}$$

Une base pour les règles d'association $\mathrm{M_{GK}}$-valides

La mesure de qualité $\mathrm{M_{GK}}$ peut s'écrire en fonction du support et de la confiance de la façon suivante :

$$M_{GK}(X \rightarrow Y) = \begin{cases} \frac{\mathrm{Conf}(X \rightarrow Y) - \mathrm{Supp}(Y)}{1 - \mathrm{Supp}(Y)} & si\ p(Y'|X') \geq p(Y') \\ \frac{\mathrm{Conf}(X \rightarrow Y) - \mathrm{Supp}(Y)}{\mathrm{Supp}(Y)} & si\ p(Y'|X') \leq p(Y'). \end{cases}$$

Cette mesure de qualité $\mathrm{M_{GK}}$ vérifie les cinq propriétés suivantes :

1. $M_{GK}(X \rightarrow Y) = -1$ si et seulement si X et Y sont incompatibles (*i.e.,* si $p(X' \cap Y') = 0$) ;
2. $-1 < M_{GK}(X \rightarrow Y) < 0$ si et seulement si X défavorise Y, ou encore si X et Y sont négativement dépendants (*i.e.,* si $p(Y'|X') < p(Y')$) ;
3. $M_{GK}(X \rightarrow Y) = 0$ si et seulement si X et Y sont indépendants (*i.e.,* si $p(Y'|X') = p(Y')$) ;
4. $0 < M_{GK}(X \rightarrow Y) < 1$ si et seulement si X favorise Y, ou encore si X et Y sont positivement dépendants (*i.e.,* si $p(Y'|X') > p(Y')$) ;
5. $M_{GK}(X \rightarrow Y) = 1$ si et seulement si X implique logiquement Y (*i.e.,* si $p(Y'|X') = 1$).

Ces propriétés expriment le fait que $\mathrm{M_{GK}}$ prend ses valeurs sur l'intervalle borné $[-1, 1]$ tout en réflétant les situations de référence telles que l'incompatibilité, la dépendance négative, l'indépendance, la dépendance positive et l'implication logique entre la prémisse et le conséquent d'une règle d'association. Ainsi, $\mathrm{M_{GK}}$ est une mesure de qualité dite normalisée au sens défini dans (Feno et al., 2006; Diatta et al., 2007) où il est également montré que $\mathrm{M_{GK}}$ est la normalisée associée à la plupart des mesures de qualité proposées dans la littérature. Elle mesure le degré de dépendance positive orientée entre la prémisse et le conséquent d'une règle d'association. En effet, pour une règle d'association $X \rightarrow Y$, $\mathrm{M_{GK}}(X \rightarrow Y)$ proche de 1 signifie que la dépendance entre X et Y est forte tandis que $\mathrm{M_{GK}}(X \rightarrow Y)$ voisine de 0 signifie que la dépendance entre X et Y est faible. Par ailleurs, le fait que $\mathrm{M_{GK}}$ est une mesure non symétrique mais implicative (*i.e.,* vérifie le principe logique de la contraposition cf. point (1) de la proposition Proposition 3) dans le cas où la prémisse d'une règle favorise son conséquent, justifie l'orientation de règles mesurées.

La proposition ci-après montre que la notion de dépendance positive (resp. négative) entre deux motifs est une relation symétrique, sans toutefois que le degré de dépendance reste constant.

Proposition 1 *Soient X et Y deux motifs. On a les équivalences suivantes*

(i) X favorise Y (i.e. $p(Y'|X') > p(Y')$) si et seulement si Y favorise X.

(ii) X défavorise Y (i.e. $p(Y'|X') < p(Y')$) si et seulement si Y défavorise X.

La proposition ci-dessous précise les liens entre les notions de dépendance positive et dépendance négative.

Proposition 2 *Soient X et Y deux motifs.*

(1) Les trois conditions suivantes sont équivalentes : (i) X défavorise Y, (ii) X favorise $\overline{Y}$ et (iii) $\overline{X}$ favorise Y.

(2) Les quatre conditions suivantes sont équivalentes : (i) X favorise Y, (ii) X défavorise $\overline{Y}$, (iii) $\overline{X}$ favorise $\overline{Y}$ et (iv) $\overline{X}$ défavorise Y.

Ces liens entre les dépendances positives et négatives montrent l'opportunité, voire la pertinence de considérer les règles négatives. En effet, lorsque X défavorise Y, il est tout a fait opportun de considérer les règles $\overline{X}\rightarrow Y$ et $X\rightarrow\overline{Y}$, puisque dans ce cas X favorise $\overline{Y}$ et $\overline{X}$ favorise Y. De même, si X favorise Y, alors il est tout à fait opportun de considérer les règles $\overline{Y}\rightarrow\overline{X}$ et $\overline{X}\rightarrow\overline{Y}$ car par dualité, dans ce cas, $\overline{X}$ et $\overline{Y}$ se favorisent mutuellement.

La proposition suivante montre, d'une part, que $\mathrm{M_{GK}}$ vérifie le principe logique de contraposition et, d'autre part, que la $\mathrm{M_{GK}}$-validité des règles négatives à droite se déduit de celle des règles négatives à gauche et vice-versa.

Proposition 3 *(1) Si X et Y sont deux motifs tels que X favorise Y, alors*

$$M_{GK}(\overline{Y}\rightarrow\overline{X}) = M_{GK}(X\rightarrow Y).$$

(2) Si X et Y sont deux motifs tels que X défavorise Y, alors

$$M_{GK}(\overline{X}\rightarrow Y) = \frac{p(X')}{1-p(X')}\frac{p(Y')}{1-p(Y')}M_{GK}(X\rightarrow\overline{Y}).$$

La poposition suivante montre que la valeur prise par $\mathrm{M_{GK}}$ sur une règle d'association positive est opposée à celle prise par $\mathrm{M_{GK}}$ appliquée à la règle négative à droite correspondante.

Proposition 4 *Soient X et Y deux motifs, nous avons la relation ci-dessous.*

$$\mathrm{M_{GK}}(X\rightarrow\overline{Y}) = -\mathrm{M_{GK}}(X\rightarrow Y)$$

Corollaire 1 *Si X et Y deux motifs tels que X défavorise Y, alors*

$$\mathrm{M_{GK}}(X\rightarrow\overline{Y}) > 0 \Leftrightarrow \mathrm{M_{GK}}(\overline{X}\rightarrow Y) > 0 \Leftrightarrow \mathrm{M_{GK}}(X\rightarrow Y) < 0.$$

Ce corollaire nous montre qu'en définitive, le cas intéressant concernera la dépendance positive entre prémisse et conséquent d'une règle d'association.

Forts de ces résultats, nous ne considérons par la suite que les règles positives et les règles négatives à droite que nous désignerons tout simplement par règles négatives (sans préciser à droite).

Soient $\alpha \in \mathbb{R}^+$ et μ une mesure de qualité des règles et X, Y deux motifs positifs ou négatifs. Alors la règle d'association $X\rightarrow Y$ sera dite (μ,α)-*valide* (ou encore μ-*valide* ou tout simplement *valide*, s'il n'y a pas un risque de confusion) si $\mu(X\rightarrow Y) \geq \alpha$.

Remarquons que $\mathrm{M_{GK}}$ peut sécrire de la manière suivante

$$\mathrm{M_{GK}}(X\rightarrow Y) = \begin{cases} \frac{p(Y'|X')-p(Y')}{1-p(Y')} & si\ X\ favorise\ Y \\ -\frac{p(\overline{Y}'|X')-p(\overline{Y}')}{1-p(\overline{Y}')} & si\ X\ favorise\ \overline{Y}. \end{cases}$$

Posons $\mathrm{M_{GK}}^f(X\rightarrow Y) = \frac{p(Y'|X')-p(Y')}{1-p(Y')}$ $si\ X\ favorise\ Y$. Alors $\mathrm{M_{GK}}$ peut s'exprimer entièrement en fonction $\mathrm{M_{GK}}^f$ de la manière suivante :

$$\mathrm{M_{GK}}(X\rightarrow Y) = 1_{(X\ favorise\ Y)}\mathrm{M_{GK}}^f(X\rightarrow Y) - 1_{(X\ \text{défavorise}\ Y)}\mathrm{M_{GK}}^f(X\rightarrow\overline{Y}),$$

où $1_{(X\ favorise\ Y)}$ désigne la fonction indicatrice de l'événement ($X\ favorise\ Y$) *i.e.*, elle a pour valeur 1 si X favorise Y et 0 sinon.

Certes, $\mathrm{M_{GK}}$ n'est pas dérivable à l'indépendance en fonction du nombre d'exemples (ou de contre-exemples), mais l'écriture ci-dessus montre que les règles négatives subissent un traitement analogue aux règles positives.

Soient X et Y de motifs (positifs et/ou négatifs). La règle d'association $X \rightarrow Y$ sera dite *exacte* si $X' \subseteq Y'$. Dans le cas contraire, $X \rightarrow Y$ sera dite *approximative*.

La proposition suivante caractérise les règles exactes (resp. approximatives) en utilisant la mesure de qualité $\mathrm{M_{GK}}$.

Proposition 5 *Soient X et Y deux motifs positifs et/ ou négatifs. La règle d'association $X \rightarrow Y$ est exacte (resp. $\mathrm{M_{GK}}$-approximative) si, et seulement si $\mathrm{M_{GK}}(X \rightarrow Y) = 1$ (resp. $0 < \mathrm{M_{GK}}(X \rightarrow Y) < 1$).*

Ainsi, ne considérant que les règles valides au sens de $\mathrm{M_{GK}}$, nous distinguerons les quatre types de règles suivantes, pour X et Y deux motifs positifs :

- les règles *positives exactes*, *i.e.*, les règles positives $X \rightarrow Y$ telles que $\mathrm{M_{GK}}(X \rightarrow Y) = 1$ égale à 1.
- les règles *négatives exactes*, *i.e.*, les règles négatives $X \rightarrow \overline{Y}$ telles que $\mathrm{M_{GK}}(X \rightarrow \overline{Y}) = 1$.
- les règles *positives* $\mathrm{M_{GK}}$*-approximatives*, *i.e.*, les règles positives $X \rightarrow Y$ telles que $\alpha \leq M_{GK}(X \rightarrow Y) < 1$.
- les règles *négatives* $\mathrm{M_{GK}}$*-approximatives*, *i.e.*, les règles négatives $X \rightarrow \overline{Y}$ telles que $\alpha \leq M_{GK}(X \rightarrow \overline{Y}) < 1$.

4 Bases pour les règles d'association $\mathrm{M_{GK}}$-valides

L'ensemble des règles d'association valides au sens d'une mesure de qualité comporte souvent un très grand nombre de règles dont plusieurs peuvent être rédondantes par rapport à des axiomes d'inférence donnés. Ainsi, d'un point de vue informatif, il est intéressant de n'en générer qu'une base, c'est-à-dire un sous-ensemble minimal à partir duquel toutes les règles valides peuvent être retrouvées par application de ces axiomes d'inférence.

Dans ce papier, nous proposons une base pour les règles d'association (M_{GK}, α)-valides où $\alpha \in]0; 1[$. Cette base est la réunion de quatre bases partielles : une base pour les règles positives exactes, une base pour les règles négatives exactes, une base pour les règles positives $\mathrm{M_{GK}}$-approximatives et une base pour les règles négatives $\mathrm{M_{GK}}$-approximatives.

4.1 Base pour les règles positives exactes

Rappelons que les règles positives $\mathrm{M_{GK}}$-exactes sont les règles $X \rightarrow Y$ telles que $M_{GK}(X \rightarrow Y) = 1$. La proposition ci-dessous montre que ces règles coïncident avec les règles positives confiance-exactes, *i.e.* les règles de confiance 1.

Proposition 6 *Soient X et Y deux motifs tels que $\mathrm{Supp}(X) \neq 0$ et $\mathrm{Supp}(Y) \neq 0$. Alors $M_{GK}(X \rightarrow Y) = 1$ si et seulement si $\mathrm{Conf}(X \rightarrow Y) = 1$.*

Il résulte de cette identité entre l'ensemble des règles positives $\mathrm{M_{GK}}$-exactes et l'ensemble des règles positives confiance-exactes que la base de Guigues-Duquenne (Guigues et Duquenne, 1986) pour les règles positives confiance-exactes est une base pour les règles positives

$\mathrm{M_{GK}}$-exactes. Rappelons que cette base est caractérisée en termes de l'opérateur $\varphi = f \circ g$, où f et g sont les applications définies par

$$\begin{array}{ll} f : \mathcal{P}(\mathcal{E}) \rightarrow \mathcal{P}(\mathcal{V}) & et \quad g : \mathcal{P}(\mathcal{V}) \rightarrow \mathcal{P}(\mathcal{E}) \\ E \longmapsto f(E) = \{x \in \mathcal{V}, \forall e \in E \; x(e) = 1\} & \quad E \longmapsto g(X) = X'. \end{array}$$

Plus précisement, cette base est l'ensemble BPE défini par

$$\mathrm{BPE} = \{X \rightarrow \varphi(X) \setminus X : X \; est \; \varphi - critique\},$$

où un motif X est φ-critique si $\varphi(X) \neq X$ et pour tout $Y \subset X$, ou bien $\varphi(Y) \subset X$ ou bien $X \subset \varphi(Y)$ et il existe $Z \subset Y$ tel que $\varphi(Z)$ intersecte proprement Y (Diatta, 2005).

Exemple 1 *Pour le contexte donné dans le tableau 1, les ensembles φ-critiques sont A, B, D, E. Ainsi,* $\mathrm{BPE} = \{A \rightarrow C, B \rightarrow E, D \rightarrow AC, E \rightarrow B\}$.

4.2 Base pour les règles négatives exactes

Rappelons que par règle négative est entendu règle négative à droite, *i.e.,* les règles $X \rightarrow \overline{Y}$ où X et Y sont des motifs positifs. La proposition ci-dessous définit leurs supports et confiances.

Proposition 7 *Soit X et Y deux motifs. Alors :*
(1) $\mathrm{Supp}(\overline{X}) = 1 - \mathrm{Supp}(X)$.
(2) $\mathrm{Supp}(X \rightarrow \overline{Y}) = \mathrm{Supp}(X) - \mathrm{Supp}(X \rightarrow Y)$.
(3) $\mathrm{Conf}(X \rightarrow \overline{Y}) = 1 - \mathrm{Conf}(X \rightarrow Y)$

Le résultat suivant montre que les règles négatives $\mathrm{M_{GK}}$-exactes coïncident avec leurs correspondantes positives de support nul.

Proposition 8 *Soient X et Y deux motifs tels* $\mathrm{Supp}(X) \neq 0$ *et* $\mathrm{Supp}(Y) \neq 0$. *Alors* $\mathrm{M_{GK}}(X \rightarrow \overline{Y}) = 1$ *si et seulement si* $\mathrm{Supp}(X \rightarrow Y) = 0$.

Le résultat de la Proposition 8 nous conduit à considérer la bordure positive de l'ensemble des motifs de support non nul, notée $\mathrm{Bd}^+(0)$ (Mannila et Toivonen, 1997) et définie par

$$\mathrm{Bd}^+(0) = \{X \subseteq \mathcal{V} : \mathrm{Supp}(X) > 0 \; et \; pour \; tout \; x \notin X, \; \mathrm{Supp}(X \cup \{x\}) = 0\}.$$

Exemple 2 *Pour le contexte donné dans le tableau 1,* $\mathrm{Bd}^+(0) = \{ACD, ABCE\}$.

Considérons maintenant les axiomes d'inférence ci-après :
(NE1) si $X \rightarrow \overline{Y}$ et T est tel que $\mathrm{Supp}(Y \cup T) > 0$, alors $X \rightarrow \overline{Y \cup T}$;
(NE2) si $X \rightarrow \overline{Y}$ et $Z \subset X$ est tel que $\mathrm{Supp}(Z \cup Y) = 0$, alors $Z \rightarrow \overline{Y}$.
Alors, on montre que les axiomes (NE1) et (NE2) sont corrects, c'est-à-dire que toute règle d'association déduite par application de (NE1) et (NE2) à partir de règles d'association négatives $\mathrm{M_{GK}}$-exactes est négative $\mathrm{M_{GK}}$-exacte. De plus, nous avons le résultat suivant (voir preuve en annexe) :

Thèoreme 1 *L'ensemble* BNE *défini par*

$$\mathrm{BNE} = \{X \rightarrow \overline{\{x\}} : x \notin X \; et \; X \in Bd^+(0)\}$$

est une base pour les règles négatives $\mathrm{M_{GK}}$*-exactes par rapport aux axiomes d'inférence (NE1) et (NE2).*

Exemple 3 *Pour le contexte donné dans le tableau 1,* $\mathrm{BNE} = \{ABCE \rightarrow \overline{D}, ACD \rightarrow \overline{B}, ACD \rightarrow \overline{E}\}$*. Par ailleurs, les* 10 *règles* $ABCE \rightarrow \overline{AD}$*,* $ABCE \rightarrow \overline{CD}$*,* $ABE \rightarrow \overline{ACD}$*,* $BE \rightarrow \overline{AD}$*,* $E \rightarrow \overline{AD}$*,* $B \rightarrow \overline{AD}$*,* $E \rightarrow \overline{CD}$*,* $B \rightarrow \overline{CD}$*,* $E \rightarrow \overline{ACD}$*,* $B \rightarrow \overline{ACD}$ *se déduisent de la règle* $ABCE \rightarrow \overline{D}$*, par application de (NE1) et (NE2).*

Remarque 1
- *La bordure positive* $\mathrm{Bd}^+(0)$ *est identique à l'ensemble des motifs* φ*-fermés maximaux de support non nul (Pasquier et al., 1999). Ainsi, la base* BNE *peut se caractériser en termes de l'opérateur de fermeture* φ*.*
- *Si la règle* $X \rightarrow \overline{Y}$ *est* $\mathrm{M_{GK}}$*-exacte alors la règle* $Y \rightarrow \overline{X}$ *l'est aussi, et réciproquement. Toutefois, les deux règles n'ont pas toujours le même degré d'informativité. En effet, si* $|X_1| > |X_2| > |Y_1| > |Y_2|$*, alors la règle* $X_2 \rightarrow \overline{Y}_2$ *est la plus informative que tout autre règle négative* $\mathrm{M_{GK}}$*-exactes combinant les motifs* X_1, X_2, Y_1, Y_2*.*

4.3 Base pour les règles positives approximatives

Notons qu'une règle valide au sens de la mesure de qualité $\mathrm{M_{GK}}$ est nécessairement une règle où la prémisse favorise la conclusion. En effet, X défavorise Y signifie que la réalisation de X diminue la chance de Y d'être réalisé. Dans ce cas il est alors plus pertinent de considérer la règle $X \rightarrow \overline{Y}$ puisque X favorise $\overline{Y}$ lorsque X défavorise Y.

Soit $\alpha \in [0, 1]$. Le résultat suivant caractérise les règles positives approximatives (M_{GK}, α)-valides en fonction de leurs confiances respectives.

Proposition 9 *Soient* X *et* Y *deux motifs tels que* X *favorise* Y*. Alors* $\alpha \leq M_{GK}(X \rightarrow Y) < 1$ *si et seulement si* $\mathrm{Supp}(Y)(1-\alpha) + \alpha \leq \mathrm{Conf}(X \rightarrow Y) < 1$.

Considérons maintenant l'axiome d'inférence (PA) ci-dessous :
(PA) si $X \rightarrow Y$ et Z, T sont tels que $\varphi(X) = \varphi(Z)$ et $\varphi(Y) = \varphi(T)$, alors $Z \rightarrow T$.
Alors on montre que (PA) est correct, c'est-à-dire que toute règle d'association déduite par application de (PA) à partir d'une règle positive $\mathrm{M_{GK}}$-approximative est positive $\mathrm{M_{GK}}$-approximative.
Par ailleurs, nous avons le résultat suivant (voir preuve en annexe) :

Thèoreme 2 *L'ensemble* $\mathrm{BPA}(\alpha)$ *défini par*

$$\mathrm{BPA}(\alpha) = \{X \rightarrow Y : \varphi(X) = X, \varphi(Y) = Y, \mathrm{Supp}(Y)(1-\alpha) + \alpha \leq \mathrm{Conf}(X \rightarrow Y) < 1\}$$

est une base pour les règles d'association positives $\mathrm{M_{GK}}$*-approximatives, par rapport à l'axiome d'inférence (PA).*

Exemple 4 *Prenons* $\alpha = \frac{1}{10}$*. Pour le contexte donné dans le tableau 1,* $AC \rightarrow BCE \in BPA(\alpha)$*. Par ailleurs, les* 5 *règles* $A \rightarrow BC, A \rightarrow CE, A \rightarrow BCE, AC \rightarrow BC, AC \rightarrow CE$ *se déduisent de la règle* $AC \rightarrow BCE$*, par application de (PA).*

4.4 Base pour les règles négatives approximatives

La proposition, ci-dessous, caractérise les règles négatives $\mathrm{M_{GK}}$-approximatives en fonction de la confiance des règles positives correspondantes.

Proposition 10 *Soient* X *et* Y *deux motifs tels que* X *défavorise* Y *i.e.* X *favorise* $\overline{Y}$*. Alors* $\alpha \leq M_{GK}(X \rightarrow \overline{Y}) < 1$ *si et seulement si* $0 < \mathrm{Conf}(X \rightarrow Y) \leq \mathrm{Supp}(Y)(1-\alpha)$.

Considérons enfin l'axiome d'inférence (NA) ci-dessous :
(NA) Si $X \rightarrow \overline{Y}$ et Z, T sont tels que $\varphi(X) = \varphi(Z)$ et $\varphi(Y) = \varphi(T)$, alors $Z \rightarrow \overline{T}$.

On montre que l' axiome (NA) est correct, c'est-à-dire que toute règle d'association déduite par application de (NA) à partir d'une règle négative $\mathrm{M_{GK}}$-approximative est négative $\mathrm{M_{GK}}$-approximative. Par ailleurs, nous avons le résultat suivant (démonstration analogue à celle du Théorème 2) :

Thèoreme 3 *L'ensemble* $\mathrm{BNA}(\alpha)$ *défini par*

$$\mathrm{BNA}(\alpha) = \{X \rightarrow \overline{Y} : \varphi(X) = X, \varphi(Y) = Y, 0 < \mathrm{Conf}(X \rightarrow Y) \leq \mathrm{Supp}(Y)(1-\alpha)\}$$

est une base pour les règles d'association négatives $\mathrm{M_{GK}}$*-approximatives, par rapport à l'axiome d'inférence (NA).*

Exemple 5 *Prenons* $\alpha = \frac{1}{10}$*. Pour le contexte donné dans le Tableau 1,*
$BNA(\alpha) = \{AC \rightarrow \overline{BE}, BE \rightarrow \overline{AC}\}$*. Par ailleurs, comme* $AC = \varphi(AC) = \varphi(A)$ *et* $BE = \varphi(BE) = \varphi(B) = \varphi(E)$*, les* 5 *règles* $A \rightarrow \overline{B}$, $A \rightarrow \overline{E}$, $A \rightarrow \overline{BE}$, $AC \rightarrow \overline{B}$, $AC \rightarrow \overline{E}$ *se déduisent de la règle* $AC \rightarrow \overline{BE}$ *par application de (NA).*

5 Conclusion

L'un des problèmes de la fouille des règles d'association est la surabondance des règles extraites à partir d'un contexte. La recherche d'une base permet de réduire de façon significative le nombre des règles générées. Nous avons présenté une base pour les règles d'association valides au sens de la mesure $\mathrm{M_{GK}}$. Cette base est la réunion de quatre bases : une base pour les règles positive exactes, une base pour les règles négatives exactes, une base pour les règles positives approximatives et une base pour les règles négatives approximatives. Nous travaillons actuellment sur des algorithmes de génération de ces bases.

Références

Agrawal, R., T. Imielinski, et A. Swami (1993). Mining association rules between sets of items in large databases. In P. Buneman et S. Jajodia (Eds.), *Proc. of the ACM SIGMOD International Conference on Management of Data*, Volume 22, pp. 207–216.

Azé, J., N. Lucas, et M. Sebag (2003). Fouille de données visuelle et analyse de facteurs de risque médical. In *EGC*, pp. 183–188.

Diatta, J. (2005). Caractérisation des ensembles critiques d'une famille de moore finie. In *Septièmes journées de la Société Francophone de Classification*, pp. 126–129.

Diatta, J., H. Ralambondrainy, et A. Totohasina (2007). Towards a unifying probabilistic implicativenormalized quality measure for association rules. *Quality Measures in Data Mining*, 237–250.

Feno, D. R., J. Diatta, et A. Totohasina (2006). Normalisée d'une mesure probabiliste de qualité des règles d'association : étude des cas. In *Actes du 2nd Qualité des Données et des Connaissances (D.K.Q.)*, pp. 25–30.

Gras, R. et A. Totohasina (1995). Chronologie et causalité, sources d'obstacles épistémologiques à l'apprentissage de probabilité conditionnelle. *Recherche en didactique des mathématiques 15*, 49–95.

Guigues, J. L. et V. Duquenne (1986). Famille non redondante d'implications informatives résultant d'un tableau de données binaires. *Mathématiques et Sciences humaines 95*, 5–18.

Guillaume, S. (2000). *Traitement des données volumineuses. Mesures et algorithmes d'extraction des règles d'association et règles ordinales*. Ph. D. thesis, Universté de Nantes, France.

Lallich, S. et O. Teytaud (2004). Evaluation et validation de mesures d'intérêt des règles d'association. *RNTI-E-1 spécial*, 193–217.

Mannila, H. et H. Toivonen (November 1997). Levelwise search and borders of theories in knowledge discovery. *Data Mining Knowledge Discovery 1. 3*, 241–258.

Pasquier, N., Y. Bastide, R. Taouil, et L. Lakhal (1999). Efficient mining of association rules using closed itemset lattices. *Information Systems 24*, 25–46.

Pei, J., J. Han, et R. Mao (2000). CLOSET : An efficient algorithm for mining frequent closed itemsets. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pp. 21–30.

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and representation of strong rules. *Knowlegde Discovery in Databases AAAI Press/The MIT Press*, 229–248.

Plott, C. (1973). Path independance, rationality and scocial choice. *Econometrica 41*, 1075–1091.

Salleb, A. (2003). *Recherche de motifs frÓquents pour l'extraction de règles d'associaition*. Ph. D. thesis, Université d'Orleans, France.

Totohasina, A. (2003). Normalisation de mesures probabilistes de la qualité des règles. In *Proc. SFDS'03, XXXV ième Journées de Statistiques*, Volume 2, pp. 985–988.

Wu, X., C. Zhang, et S. Zhang (2004). Efficient mining of both positive and negative association rules. *ACM Transactions on information Systems 3*, 381–405.

Annexe

Preuve du Théorème 1 : Nous commençons par montrer que toute règle négative $\mathrm{M_{GK}}$-exacte peut être dérivée de BNE par application de (NE1) et/ou (NE2). Soit $X \rightarrow \overline{Y}$ une règle négative $\mathrm{M_{GK}}$-exacte. Alors $\mathrm{Supp}(X) \neq 0$ *et* $\mathrm{Supp}(X \cup Y) = 0$. Ainsi, d'une part, il existe $Z \in Bd^{+}(0)$ tel que $X \subseteq Z$. D'autre part, il existe $x \in Y$ tel que $x \notin Z$ car $\mathrm{Supp}(Z) \neq 0$, $X \subseteq Z$ et $\mathrm{Supp}(X \cup Y) = 0$. Ainsi, la règle $Z \rightarrow \overline{x}$ appartient BNE. Donc, l'application de (NE1) à $Z \rightarrow \overline{x}$ donne la règle $Z \rightarrow \overline{\{x\} \cup Y}$, *i.e.*, la règle $Z \rightarrow \overline{Y}$. En outre, l'application de (NE2) à $Z \rightarrow \overline{Y}$ donne la règle $X \rightarrow \overline{Y}$ car $X \subseteq Z$ et $\mathrm{Supp}(X \cup Y) = 0$.

Montrons maintenant que l'ensemble BNE est minimal. Soit $X \rightarrow \overline{x}$ un élément de BNE et soit $\mathrm{BNE}' = \mathrm{BNE} - \{X \rightarrow \overline{x}\}$. Montrons que la règle $X \rightarrow \overline{x}$ ne peut pas être dérivée de BNE' par application de (NE1) et (NE2). En effet, la règle $X \rightarrow \overline{x}$ ne peut pas être dérivée d'une règle $X \rightarrow \overline{Y}$ par application de (NE1) car cela impliquerait nécessairement $Y \subset \{x\}$. D'autre part, la règle $X \rightarrow \overline{x}$ ne peut pas être dérivée d'une autre règle $Z \rightarrow \overline{x}$ par application (NE2). En effet,

cela impliquerait que $X \subset Z$ donc $\mathrm{Supp}(Z) = 0$ puisque $X \in Bd^+(0)$. D'où, la règle $X \rightarrow \overline{x}$ ne peut pas être dérivée d'une règle de BNE', ce qui démontre la minimalité de BNE. □

Les deux lemmes suivants seront utiles pour la démonstration des Théorème 2 et 3. Le Lemme 1 montre que le support d'un motif est égal au support de sa fermeture (Pasquier et al., 1999).

Lemme 1 *Pour tout motif* X, $\mathrm{Supp}(\varphi(X)) = \mathrm{Supp}(X)$.

Le Lemme 2 est une caractérisation des opérateurs de fermeture utilisant une propriété dite d'indépendance de chemins (Plott, 1973).

Lemme 2 *Une application extensive* ϕ *sur* $\mathcal{P}(\mathcal{V})$ *, i.e.* $X \subseteq \phi(X)$*, est un opérateur de fermeture sur* $\mathcal{P}(\mathcal{V})$ *si et seulement si elle vérifie la propriété* $\phi(X \cup Y) = \phi(\phi(X) \cup \phi(Y))$*, pour tous* $X, Y \in \mathcal{P}(\mathcal{V})$.

Preuve du Théorème 2 : Nous commençons par montrer que toute règle positive $\mathrm{M_{GK}}$-approximative peut être dérivée de $\mathrm{BPA}(\alpha)$ par application de l'axiome (PA). Soit $X \rightarrow Y$ une règle positive $\mathrm{M_{GK}}$-approximative. Alors, par la Proposition 9, $\mathrm{Supp}(Y)(1-\alpha)+\alpha \leq \mathrm{Conf}(X \rightarrow Y) < 1$. Considérons les deux motifs φ-fermés $Z = \varphi(X)$ et $T = \varphi(Y)$. D'une part, par le Lemme 1, $\mathrm{Conf}(\varphi(X) \rightarrow \varphi(Y)) = \mathrm{Supp}(\varphi(X) \cup \varphi(Y))/\mathrm{Supp}(\varphi(X))$= $\mathrm{Supp}(\varphi(\varphi(X) \cup \varphi(Y)))/\mathrm{Supp}(\varphi(X))$ qui, par le Lemme 2, est égale à $\mathrm{Supp}(\varphi(X \cup Y))/\varphi(X)$ et qui, encore par le Lemme 1, est égale à $\mathrm{Supp}(X \cup Y)/sup(X) = \mathrm{Conf}(X \rightarrow Y)$. D'autre part, par le Lemme 1, $\mathrm{Supp}(\varphi(Y) = \mathrm{Supp}(Y))$, donc $\mathrm{Supp}(\varphi(Y))(1-\alpha)+\alpha) \leq \mathrm{Conf}(\varphi(X) \rightarrow \varphi(Y)) < 1$. Donc, par la Proposition 9, $0 < \mathrm{M_{GK}}(\varphi(X) \rightarrow \varphi(Y)) < 1$ alors $\varphi(X) \rightarrow \varphi(Y)$ est un élément de $\mathrm{BPA}(\alpha)$. Par ailleurs, l'application de (PA) à $Z \rightarrow T$ donne la règle $X \rightarrow Y$.

Montrons maintenant que $\mathrm{BPA}(\alpha)$ est minimal. Soit $X \rightarrow Y$ un élément de $\mathrm{BPA}(\alpha)$ et soit $\mathrm{BPA}'(\alpha) = \mathrm{BPA}(\alpha) - \{X \rightarrow Y\}$. Nous montrons que la règle $X \rightarrow Y$ ne peut pas être dérivée de $\mathrm{BPA}'(\alpha)$ par application (PA). En effet, si $X \rightarrow Y$ pouvait être dérivée de $BPA'(\alpha)$, alors, il existerait une suite finie de règles $X_1 \rightarrow Y_1, \cdots, X_n \rightarrow Y_n$ ($n > 1$) telle que :

- $X_1 \rightarrow Y_1 \in \mathrm{BPA}'$;
- $X_n \rightarrow Y_n = X \rightarrow Y$;
- pour $i = 1, \cdots, n-1 : \varphi(X_i) = \varphi(X_{i+1})$ et $\varphi(Y_i) = \varphi(Y_{i+1})$.

Alors $X_1 = \varphi(X_1) = \cdots = \varphi(X_n) = \varphi(X) = X$ et $Y_1 = \varphi(Y_1) = \cdots = \varphi(Y_n) = \varphi(Y) = Y$ avec $X_1 \rightarrow Y_1 \in \mathrm{BPA}'$, ce qui contredit le fait que $X \rightarrow Y \notin \mathrm{BPA}'(\alpha)$. Donc, $X \rightarrow Y$ ne peut pas être dérivée de $BPA'(\alpha)$, démontrant la minimalité de $\mathrm{BPA}(\alpha)$. □

Summary

This paper is concerned with the association rules valid under the quality measure $\mathrm{M_{GK}}$. The quality measure $\mathrm{M_{GK}}$ is known to be normalized in the sense that its values lie in the interval $[-1, +1]$ and reflect the reference situations such as incompatibility, negative dependance, independance, positive dependance and logical implication between the premise and the consequent of a rule. Moreover, its properties allow to restrict oneself in considering only positive association rules and riht-hand side negative ones. Thus, we propose a basis for positive exact rules, a basis for negative exact rules, a basis for positive approximate rules and a basis for

negative approximate rules. The union of these four bases forms a basis for all positive and negative association rules valid under the quality measure $\mathrm{M_{GK}}$.