Thèse

Présentée à L'Université de La Réunion

Par

Daniel Rajaonasy **FENO**

en vue de l'obtention du

Doctorat de l'Université de La Réunion

Spécialité : **Mathématiques et Informatique**

Mesures de qualité des règles d'association : normalisation et caractérisation des bases

Soutenue le 1er décembre 2007 devant la commission d'examen :

M. **Henri RALAMBONDRAINY** Professeur (Université de La Réunion, France) Président

M. **Engelbert MEPHU NGUIFO** Maître de Conférences HDR(Université de Lens France) Rapporteur

M. **Amedeo NAPOLI** Directeur de Recherche (CNRS LORIA, France) Rapporteur

M. **Sadock BEN YAHIA** Maître Assistant(Université de Tunis El-Manar, Tunisie) Examinateur

M. **Jean DIATTA** Professeur (Université de La Réunion France) Directeur

M. **André TOTOHASINA** Maître de Conférences(Université d'Antsiranana Madagascar) Co-Directeur

Institut de **RE**cherche en **M**athématiques et **I**nformatique **A**ppliquées (**IREMIA)**

Extraction de **C**onnaissances à partir de **D**onnées **(ECD)**

REMERCIEMENTS

Cette thèse a été financée par une bourse en alternance de l'**A**gence **U**niversitaire de la **F**rancophonie (A.U.F.) pour la période de fin 2004 à 2007, soit 30 mois sur 36. Je tiens à remercier l'A.U.F. pour ses soutiens financiers pendant mes trois années de thèse.

Je ne pense pas que quelques mots de remerciements puissent suffire pour exprimer le sentiment de profonde gratitude et de reconnaissance que j'éprouve à l'endroit de mes Directeurs de thèse Monsieur Jean DIATTA, Professeur à l'Université de La Réunion (France) et Monsieur André TOTOHASINA, Maître de Conférences à l'Université d'Antsiranana, Madagascar, pour m'avoir encadré avec diligence, disponibilité totale et une clairvoyance remarquable pour ces travaux. Qu'ils trouvent ici mes remerciements les plus sincères.

Je remercie Monsieur Amedeo NAPOLI, Directeur de Recherche CNRS, au LORIA, Nancy, France, et Monsieur Engelbert MEPHU GUIFO, Maître de Conférences HDR à l'IUT de Lens, France, d'avoir accepté d'être les rapporteurs de ma thèse. Je les remercie pour l'attention avec laquelle ils ont lu et évalué ce mémoire.

Je remercie également Monsieur Henri RALAMBONDRAINY, Professeur à l'Université de La Réunion, France, pour l'honneur qu'il me fait d'avoir accepté d'être Président du jury.

Je remercie Monsieur Sadok BEN YAHIA, Maître Assistant à l'Université de Tunis, Tunisie, d'avoir accepté d'être membre du jury.

Je remercie également toute l'équipe du laboratoire IREMIA de l'Université de La Réunion, pour l'accueil, en particulier Monsieur Jean Guy AVELIN, Ingénieur de Recherche et Monsieur Fanilo HARIVELO, jeune Docteur en Informatique, pour leurs soutiens constants et leurs aides.

Je remercie mille fois mon épouse et toute ma famille pour leurs amours et leurs soutiens moraux.

Mes remerciements vont aussi à tous mes amis, étudiants et étudiantes Malagasy à l'Université de La Réunion, pour l'ambiance véritablement chaleureuse dans laquelle j'ai vécu durant tous mes séjours à l'Université de La Réunion.

TABLE DES MATIÈRES

1. INTRODUCTION

L'Extraction de Connaissances à partir (des Bases) de Donnes (E.C.D. parfois notée E.C.B.D.) est une discipline récente qui recoupe les domaines des bases de données, des statistiques, de l'intelligence artificielle et de l'interface homme/machine. Le défi de l'Extraction de Connaissances à partir des Données consiste à exploiter automatiquement des informations généralisables en connaissances nouvelles sous le contrôle des experts des données[Bri04, BC07, LHCM00]. Cela nécessite la conception et la mise au point de méthodes pour extraire les informations et les transformées en connaissance apportant une plus-value aux experts. Les systèmes de l'E.C.D. incorporent des théories, des algorithmes et des méthodes des différents domaines qui recoupent l'E.C.D.

Un processus de l'E.C.D. met en jeu, de manière interactive et itérative, de multiples méthodes pour la préparation des données, leur exploration, la visualisation et l'interprétation [FPSS96]. D'une part, la nature itérative du processus se justifie par son fonctionnement où les expériences s'effectuent sous forme d'essai-erreur et sont repétées en boucle. Cela consiste à : expliciter les connaissances expertes, identifier les concepts et relations entre concepts du domaine, établir la correspondance entre les concepts et les données brutes et résoudre les ambiguïtés, organiser convenablement les données, conserver la trace des expériences précédentes Les méthodes de l'E.C.D. proposent des solutions aux problèmes de recherche des règles d'association, de classification supervisée et non supervisée. Une étape centrale de ces processus est la découverte des motifs telles que les régularités, les rgles, les concepts, etc. Ces derniers capturent des informations spécifiques de la base de données. Compte tenu des tailles des bases de données (qui comprennent souvent des milliers d'attributs décrits par des millions d'entités), il s'agit d'un problème algorithmiquement ardu nécessitant la conception de méthodes efficaces pour parcourir l'espace de recherche, pour permettre la réduction du nombre de motifs ainsi extraits sans perte de l'information. Les travaux de recherche effectués dans le présent travail s'inscrivent dans le cadre de l'Extraction de

Connaissances à partir de Données et plus particulièrement dans le domaine de la fouille des règles d'association.

Pour filtrer les règles d'association intéressantes d'un contexte de la fouille de données binaires, on utilise des critères communément appelés mesures de qualité de règles. Certes la qualité d'une règle doit être objectivement évaluée, mais notamment en vue de l'interprétation, elle possède aussi un caractère subjectif en tenant compte des connaissances ou convictions a priori des experts [LHCM00, OKO04]

Plusieurs mesures de qualité [GH06] ont été proposées dans la littérature (environ une quarantaine). Ce nombre important de mesures de qualité engendre de nouveaux problèmes, entre autres, le choix de mesures de qualité à utiliser pour la fouille des règles d'association. La littérature atteste que plusieurs travaux ont été effectués pour aider les utilisateurs ou experts dans le choix de mesure de qualité [LT04, LFZ99, LMVP03, BGBG05]. Dans le présent travail, nous proposons une approche analytique appelée *normalisation* d'une mesure de qualité permettant d'apporter un nouvel éclairage sur l'étude des mesures de qualité des règles d'association. Nous montrerons que la plupart des mesures de qualité proposées dans la littérature ont une normalisée cummune [FDT06a] à savoir la mesure M_{GK} introduite indépendamment dans [Gui00] et dans [WZZ04]. Profitant des résultats obtenus par le processus de normalisation de mesures de qualité des règles, nous caractérisons une base qui est composée de quatre sous-bases pour les règles d'association valides au sens de la mesure de qualité M_{GK} [FDT06b, DFT06]. Notons que l'extraction de bases dans la fouille des règles d'association permet de réduire, sans perte d'information, le nombre de règles à présenter aux utilisateurs.

Ce mémoire se décompose en huit chapitres. Le deuxième chapitre concerne les notions utiles considérées comme les fondements mathématiques de la notion de la fouille des règles d'association. Il concerne les états de l'art sur les notions des treillis, les familles de Moore et les correspondances de Galois.

Le troisième chapitre concerne les méthodes de l'extraction des règles d'association d'un contexte de la fouille de données binaires. Généralement, l'extraction des règles d'association d'un contexte de la fouille de données se subdivise en deux sous-problèmes à savoir

1- Extraction des motifs fréquents ; en fait, on peut se contenter d'extraire les motifs fermés fréquents [Zak00a] et les motifs fréquents en serons

ensuite dérivés

2- Dérivation des règles d'association valides à partir des motifs fréquents déduits des motifs fermés fréquents [Zak00a].

La résolution du second sous-problème est aisée, elle ne nécessite pas les parcours coûteux de la base de données. Plus d'efforts ont été consacrés au premier sous-problème. Nous développons, dans ce chapitre, quelques méthodes d'extraction des représentations condensées des motifs fréquents (*i.e.*, un sous-ensemble des motifs les fréquents à partir duquel, on peut retrouver tous les motifs fréquents).

Dans le quatrième chapitre, nous présentons des études sur les différentes mesures de qualité des règles d'association. Notons que les mesures de qualité sont naturellement utilisées pour capturer les règles utiles et pertinentes à partir d'un contexte de la fouille de données. Nous présentons des critères d'appréciation de ces mesures. Nous pouvons constater que parmi les mesures disponibles dans la littérature, aucune d'elles ne satisfait à l'ensemble de ces critères. Par ailleurs, la mesure de qualité Confiance [AIS93] souvent utilisée par les méthodes de l'extraction des règles d'association suscite plusieurs critiques. Eu égard à ces problématiques, nous étudions de façon approfondie les comportements des mesures de qualité. Grâce à ses riches et intéressantes propriétés mathématiques développées par la suite, notre choix tombe sur la mesure de qualité M_{GK} [Gui00] qui tient compte de plusieurs situations de référence telles que l'incompatibilité, la répulsion, l'indépendance, l'attraction et l'implication entre les motifs d'une règle d'association.

Le cinquième chapitre est essentiellement consacré aux études des propriétés mathématiques de la mesure de qualité M_{GK}. Signalons que cette mesure s'avère ainsi intéressante, car elle répond à une grande partie des souhaits des chercheurs [PS91, LT04, BGBG05, Fre99] travaillant dans ce domaine de la fouille des règles d'association. Ensuite, elle apparaît bien adaptée à l'extraction simultanée des règles d'association dites négatives (*i.e.*, des règles présentant de négations de motifs),. et celles positives. Nous présentons dans le sixième chapitre la normalisation d'une mesure de qualité des règles d'association. Cette approche analytique, nous permet de classifier les mesures de qualité proposées dans la littérature selon trois classes :

(a) mesures de qualité M_{GK}-normalisables,

(b) mesures de qualité normalisables dont la normalisée est différente de M_{GK},

(c) mesures de qualité non normalisables.

Notons que la première classe contient la plupart des mesures de qualité disponibles dans la littérature.

Le septième chapitre concerne des caractérisations de bases, (*i.e.*, un ensemble minimal des règles valides à partir duquel, on peut dériver toutes les règles valides par utilisation d'un ensemble d'axiomes d'inférence) pour les règles d'association valides au sens d'une mesure de qualité des règles. En profitant des résultats du sixième chapitre, nous caractérisons une base au sens de la mesure de qualité M_{GK}.

Nous terminons dans le huitième chapitre par une conclusion générale en soulevant quelques questions qui nous semblent requérir davantage d'attention.

2. PRÉLIMINAIRES

2.1 Introduction

Nous présentons dans ce chapitre des notions mathématiques, à savoir des Treillis, des Familles de Moore et des Correspondances de Galois, qui seront utiles dans la théorie de la fouille des règles d'association. Le treillis est une notion très ancienne qui trouve son origine vers la moitié de dix-neuvième siècle issue des travaux de Boole [Boo47], qui s'intitule *The Mathematical Analysis of Logic*. La théorie de treillis a été longtemps négligée, avant de prendre son évolution à partir de l'année 1940 par les travaux de Öre [Ö44], BirKhoff [Bir67] la transformant en branche fertile de l'Algèbre. Elle est présente dans plusieurs domaines de mathématiques pures et appliquées, en particulier dans la fouille des règles d'association. La fouille des règles d'association consiste à trouver les liens pertinents et non triviaux entre motifs (ensembles d'attributs) dans une grande base de données. L'ensemble de tous les motifs d'une grande base de données a la structure de treillis. Comme la notion de treillis, les familles de Moore, dites aussi systèmes de fermeture, se sont avérées fondamentales pour plusieurs domaines de l'informatique (base de données, analyses formelles des concepts). Cette notion de familles de Moore est connue depuis Armstrong [Arm74] : se donner une famille de Moore sur un ensemble fini E équivaut à se donner un système d'implications entre parties de E. Les correspondances de Galois jouent un rôle important dans la théorie des ensembles ordonnés depuis leur mise en évidence par Birkhoff en 1940 sous le nom d'applications polarisées (cf. [Bir67]). Everett [Eve55] a montré que toute fermeture peut être considérée comme résultante d'une correspondance de Galois.

Ce chapitre est organisé de la façon suivante. Dans la Section 2.2, nous rappelons quelques notions fondamentales sur les treillis. La Section 2.3 concerne la notion de Familles de Moore et ses notions équivalentes. Dans la Section 2.4, nous parlons des correspondances de Galois à partir desquelles sont obtenues des treillis particuliers appelés *treillis de Galois* ou *treillis de*

concepts formels qui servent comme fondement mathématiques de la fouille des règles d'association.

2.2 Treillis

2.2.1 Ensembles ordonnés

Définition 1: Soient E et F deux ensembles.

- Une *relation binaire* de E vers F, notée $\mathcal{R}$, est un sous-ensemble de $E \times F$. On écrit $x\mathcal{R}y$ ou $(x, y) \in \mathcal{R}$ pour signifier que x est en relation avec y par $\mathcal{R}$. Dans le cas où $F = E$, on dit que $\mathcal{R}$ est une relation binaire sur E.

- Une relation binaire $\mathcal{R}$ sur E est dite *relation d'ordre (partiel)* si elle vérifie les trois conditions suivantes :

  (i) pour tout $x \in E, x\mathcal{R}x$ (réflexivité) ;

  (ii) pour tous $x, y, z \in E$, $x\mathcal{R}y$ et $y\mathcal{R}z$ impliquent $x\mathcal{R}z$ (transitivité) ;

  (iii) pour tous $x, y \in E$, $x\mathcal{R}y$ et $y\mathcal{R}x$ impliquent $x = y$ (antisymétrie).

- Un *ensemble ordonné* est un couple $(E, \mathcal{R})$ où E est un ensemble et $\mathcal{R}$ une relation d'ordre définie sur E. On note plus souvent par "$\leq$" une relation d'ordre $\mathcal{R}$. Par abus de langage, on dit alors que l'ensemble E est un ensemble ordonné (au lieu de $(E, \mathcal{R})$ ou $(E, \leq)$).

- Soient $x, y \in E$: x et y sont dits *comparables* si, et seulement si ou bien $x \leq y$, ou bien $y \leq x$. Si $x \leq y$ et $x \neq y$, on écrit $x < y$. Dans le cas contraire, x et y sont dits *incomparables*.

- On dit qu'une relation ordre $\leq$ est un ordre *total*, si pour tous $x, y \in E$, x et y sont comparables. Dans ce cas, $(E, \leq)$ est dit ensemble *totalement ordonné*.

- Une partie A d'un ensemble ordonné E est dit *chaîne*, si tous les éléments de A sont comparables. Une partie A de E est dit *anti-chaîne*, si deux éléments distincts de A ne sont jamais comparables.

- La *longueur* d'un ensemble ordonné est la taille maximale d'une chaîne moins un de cet ensemble. La *largeur* d'un ensemble ordonné est la taille maximale d'une anti-chaîne de cet ensemble.

Exemple 1: Considérons l'ensemble $E = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ et notons "$|$" la relation de divisibilité sur E : $\forall x, y \in E, x|y \Leftrightarrow (\exists q \in N : y = qx.)$. Le couple $(E, |)$ est un ensemble ordonné. Les ensembles : $\{1, 2, 4, 8, 0\}, \{1, 2, 6, 0\}$, $\{1, 3, 6, 0\}, \{1, 3, 9, 0\}, \{1, 5, 0\}$, $\{1, 7, 0\}$ sont des chaînes de E. La longueur de l'ensemble ordonné $(E, |)$ (*i.e.* la taille maximale d'une chaîne de E moins un) est égale à 4.

Les ensembles : $\{2, 3, 5, 7\}, \{4, 6, 9\}, \{8, 6, 9, 7\}, \{8, 6, 9, 5, 7\}$ sont des anti-chaînes de E. La largeur de l'ensemble ordonné $(E, |)$ (*i.e.*, la taille maximale d'une anti-chaîne de E) est égale à 5.

Définition 2: Soient E un ensemble ordonné et, x et y deux éléments de E. On dit que y *couvre* x ou x *est couvert* par y, et on note $x \prec y$, si $x < y$ et pour tout $z \in E$ tel que $x \leq z < y$ alors $z = x$, *i.e.*, il n'existe pas $z \in E$ tel que $x < z < y$.

Remarque 1: Soit E un ensemble ordonné et fini. Lorsqu'une relation de couverture existe sur E, l'ordre est représentable, à travers cette relation de couverture, par un graphe appelé *diagramme de Hasse*. Les éléments x et y sont respectivement représentés par des points p_x et p_y. La construction du diagramme de Hasse se base sur les deux principes suivants :

- si $x < y$ alors p_x est en dessous de p_y ;
- p_x et p_y sont liés par un segment si, et seulement si $x \prec y$.

Définition 3: Soient $(E, \leq)$ un ensemble ordonné et A une partie de E. La restriction de la relation d'ordre $\leq$, à l'ensemble A, notée $\leq_A$, est une relation d'ordre. On dit alors que $(A, \leq_A)$ est un *sous-ensemble ordonné* de $(E, \leq)$ et on le notera aussi $(A, \leq)$.

Définition 4: (Dual d'un ensemble ordonné)
Étant donné un ensemble ordonné $(E, \leq)$, on peut créer un nouvel ensemble ordonné, noté $(E, \leq^d)$, appelé le *dual* de E. La relation duale $\leq^d$, notée aussi $\geq$, est définie par, pour tous $x, y \in E$, $x \leq^d y$ si, et seulement si $y \leq x$.

Théorème 1: (Principe de dualité)[BM70]
Étant donnée une proposition ϕ vraie dans tout ensemble ordonné, alors la proposition duale ϕ^d est aussi vraie dans tout ensemble ordonné.

Définition 5: Soient $(E, \leq)$ un ensemble ordonné, A une partie de E et x un élément de A. On dit que :

- x est un élément *maximal* dans A si pour tout $y \in A$, $x \leq y$ implique $x = y$.
- x est un élément *minimal* dans A si pour tout $y \in A$, $y \leq x$ implique $x = y$.
- x est l'élement *maximum* de A si pour tout $y \in A$, $y \leq x$ (dans ce cas, l'élément x est unique).
- x est le *minimum* de A si pour tout $y \in A$, $x \leq y$ (dans ce cas, l'élément x est unique).
- Si A est une partie non vide de E, un élément x de E est un *majorant* (resp. *minorant*) de A si pour tout $a \in A$, $a \leq x$ (resp. $x \leq a$). On notera $MajA$ (resp. $MinA$) l'ensemble des majorants (resp. minorants) de A. On pose que $Maj\emptyset = Min\emptyset = E$.
- L'élément minimum de $MajA$ (s'il existe) est appelé la *borne supérieure* ou le *supremum* de A et on le note $\bigvee A$ *ou* $sup(A)$. Dualement, l'élément maximum de $MinA$ (s'il existe) est appelé la *b*orne inférireure ou l'*infimum* de A et on le note $\bigwedge A$ *ou* $inf(A)$.
- Une partie A de E est dite *bornée*, si elle admet à la fois une borne supérieure et une borne inférieure.

Remarque 2: Si $A = \{x, y\}$, on note $x \vee y$ (resp. $x \wedge y$) le supremum (resp. l'infimum) de A (s'il existe). Inversement, l'ordre peut s'écrire en terme de supremun et d'infimum, car pour tous $x, y \in E$, on a $x \leq y \Leftrightarrow x \vee y = y \Leftrightarrow x \wedge y = x$.

2.2.2 Structures de treillis

Dans ce paragraphe, nous présentons quelques généralités sur la notion de treillis. Dans la littérature, il existe deux définitions équivalentes de treillis : une concernant la relation d'ordre et l'autre algébrique. La Définition 6 ci-dessous est la définition algébrique du treillis proposée par Dedekind [Ded00].

Définition 6: Un treillis est un ensemble T muni de deux lois internes habituellement notées $\vee$ et $\wedge$, pour tous x, y et z, vérifiant :

$x \wedge (y \wedge z) = (x \wedge y) \wedge z$	$x \vee (y \vee z) = (x \vee y) \vee z$	(associativité)
$x \wedge y = y \wedge x$	$x \vee y = y \vee x$	(commutativité)
$x \wedge x = x$	$x \vee x = x$	(idempotence)
$x \wedge (y \vee x) = x$	$x \vee (y \wedge z) = x$	(absorption).

Nous donnons ci-après une autre définition de treillis en terme de l'ensemble ordonné.

Définition 7: Considérons un ensemble ordonné T.

- On dit que T est un *inf-demi-treillis* (resp. *sup-demi-treillis*) si pour tout couple $(x, y) \in T \times T$ l'infimum $x \wedge y$ (resp. le supremum $x \vee y$) existe.
- Un *treillis* est un ensemble ordonné qui est à la fois inf-demi-treillis et sup-demi-treillis.

Remarque 3: Les deux définitions ci-dessus sont équivalentes. En effet, si $(T, \leq)$ est un treillis selon la Définition 7, il est facile de vérifier que les deux opérations $sup(x, y) = x \vee y$ et $x \wedge y$ satisfont aux huit conditions de la Définition 6. Réciproquement, si $(T, \wedge, \vee)$ est un treillis selon la Définition 6, on peut montrer sans difficulté que la relation définie sur T par $x \leq y$ si $x \wedge y = x$ (ou, équivalente à $x \vee y = y$) est une relation d'ordre sur T, donc, l'ordre $\leq$ ainsi défini vérifie les conditions de la Définition 7.

Exemple 2:
- L'ensemble des entiers naturels muni de son ordre usuel est un treillis.
- Soit E un ensemble ordonné. L'ensemble $\mathcal{P}(E)$ de parties de E, muni de la relation d'inclusion, c'est à dire $(\mathcal{P}(E), \subseteq)$, est un treillis appelé *treillis booléen* avec, pour tous $A, B \subseteq E$, $A \vee B = A \cup B$, $A \wedge B = A \cap B$.

Théorème 2: Un sup-demi-treillis (resp. inf-demi-treillis) ayant un plus petit élément (resp. plus grand) est un treillis [Mon03].

Définition 8: On appelle *sous-treillis* d'un treillis T tout sous-ensemble non vide Q de T tel que pour tous $a, b \in Q$ $a \vee b$ *et* $a \wedge b$ appartiennent à Q.

Définition 9: Un inf-demi-treillis (resp. sup-demi-treillis) T est dit *complet* si pour toute partie non vide X de T, l'infimum $\bigwedge X$ (resp. le supremum $\bigvee X$) existe. Un treillis T est dit *complet* s'il est à la fois sup-demi-treillis complet et inf-demi-treillis complet. Tout treillis complet a un plus petit élément $\bigwedge T$, noté 0_T, et un plus grand élément $\bigvee T$, noté 1_T.

Remarque 4: Dans un treillis fini, on peut toujours définir une relation de couverture. Donc, tout treillis fini peut être représenté par un diagramme de Hasse.

Définition 10:

- Un élément x de T est dit *sup-réductible*, si il existe dans T des éléments $x_1,\ x_2$ tels que $x = x_1 \vee x_2$ avec $x_1 < x$ *et* $x_2 < x$. Un élément x de T ne possédant pas de décomposition de cette forme est dit *sup-irréductible*. On note J_T (*ou* J s'il n'y a pas de confusion) l'ensemble des éléments sup-irréductibles d'un treillis T.

- Dualement, un élément x *de* T est dit *inf-réductible* s'il existe $x_1,\ x_2 \in T$ tels que $x = x_1 \wedge x_2$, avec $x_1 > x$ *et* $x_2 > x$. Un élément x de T ne possédant pas de décomposition de cette forme est dit *inf-irréductible*. On note M_T (*ou* M[1] s'il n'y a pas risque de confusion) l'ensemble des éléments inf-irréductibles d'un treillis T.

Proposition 1: Soit T un treillis fini.

- Un élément $j \in T$ est un sup-irréductible, si et seulement si "j" couvre un seul élément.

- Un élément $m \in T$ est inf-irréductible, si et seulement si "m" est couvert par un seul élément.

[1] les notations J *et* M viennent de la terminologie anglo-saxonne, où, J (supremum) pour join et M pour meet (infimum)

Remarque 5: On peut déduire de la Proposition 1 que dans un treillis fini (de cardinalité ≥ 2), on a toujours des éléments irréductibles, par exemple, les éléments couvrant le minimum appelés *atomes*, les éléments couverts par le maximum appelés *coatomes* d'un treillis fini.

Définition 11: (1) Un treillis T est dit *atomistique* (resp. *coatomistique*), si tous les éléments sup-irréductibles (resp. inf-irréductibles) sont atomes (resp. coatomes).

(2) Un treillis T est dit *modulaire*, si pour tous $x, y, z \in T$ tels que $x \leq z$, on a $x \vee (y \wedge z) = (x \vee y) \wedge z$.

(3) Un treillis T est dit *distributif* si pour tous $x, y, z \in T, x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge y)$ (ou, équivalente à $x \vee (y \wedge z) = (x \vee y) \wedge (x \vee y)$)

(4) Un treillis *Booléen* est un treillis T qui est à la fois distributif et atomistique.

Remarque 6: Les éléments irréductibles d'un treillis permettent d'obtenir une représentation condensée de treillis tout en conservant les informations contenues dans le treillis initial. La recherche de représentation condensée concerne différents domaines informatiques parmi lesquels ceux mentionnés dans[Dom02]) : classification conceptuelle, apprentissage, traitement d'image, etc., où leur taille potentiellement exponentielle devient problématique.

2.3 Familles de Moore et Notions équivalentes

Des études menées sur les Familles de Moore appelée aussi Systèmes de Fermeture permettent d'obtenir qu'elles sont criptomorphes à d'autres notions telles que : les Opérateurs de Fermeture, Systèmes implicatifs, etc. [Dom02, DL04]. Cette section est consacrée aux études de correspondances réciproques entre la notion de Familles de Moore et Opérateurs de Fermetures, Familles de Moore et Systèmes Implicatifs.

2.3.1 Généralités sur la notion de familles de Moore

Définition 12: Soit E un ensemble. Une *famille de Moore* sur E est une partie $\mathcal{F}$ de l'ensemble $\mathcal{P}(E)$ de parties de E vérifiant les deux conditions suivantes:

(M1) $E \in \mathcal{F}$;

(M2) $\mathcal{F}' \subseteq \mathcal{F}$ implique $\cap \mathcal{F}' \in \mathcal{F}$.

Si $\mathcal{F}$ est un ensemble fini, donc E l'est aussi, la condition (M2) peut être remplacée par la condition (M'2) suivante :

(M'2) $F_1, F_2 \in \mathcal{F}$ impliquent $F_1 \cap F_2 \in \mathcal{F}$.

Les éléments de $\mathcal{F}$ sont appelés les *fermés* de $\mathcal{F}$.

Dans la suite, nous ne considérerons que les familles de Moore finies.

Exemple 3: Soit $E = \{a, b, c, d, e\}$. Alors $\mathcal{F} = \{\emptyset, a, b, d, de, bcd, abcde\}$ est une famille de Moore sur l'ensemble E. Ici, les ensembles finis sont notés comme des mots. Par exemple "ae" désigne la paire $\{a, e\}$.

Exemple 4: Soient E un ensemble et, A *et* B deux parties de E. Alors, la famille $\mathcal{F}_{A,B}$, de sous-ensembles de E, définie par : $\mathcal{F}_{A,B} = \{X \subseteq E : A \nsubseteq X$ *ou* $B \subseteq X\}$, est une famille de Moore. En particulier :

- pour $A = \emptyset$, $\mathcal{F}_{\emptyset,B} = \{X \subseteq E : B \subseteq X\}$
- Pour $B = \{i\}, \mathcal{F}_{A,\{i\}} = \{X \subseteq E : A \nsubseteq X$ *ou* $i \in X\}$ est noté $\mathcal{F}_{A,i}$.
  En particulier, $\mathcal{F}_{\{j\},i} = \{X \subseteq E : j \notin X$ *ou* $i \in X\}$ est noté $\mathcal{F}_{j,i}$.

Remarque 7: Notons que $\mathcal{F}_{A,B} = \mathcal{F}_{A,B-A}$. En effet, si $X \in \mathcal{F}_{A,B}$ alors il est clair que $X \in \mathcal{F}_{A,B-A}$. Réciproquement, si $X \notin \mathcal{F}_{A,B}$, *i.e.* $A \subseteq X$ *et* $B \nsubseteq X$, alors $A \subseteq X$ *et* $B - A \nsubseteq X$ *i.e.*, $X \notin \mathcal{F}_{A,B-A}$. Par ailleurs, d'après la définition de $\mathcal{F}_{A,B}$, si F est un fermé qui contient A alors F contient aussi B, en d'autres termes A *implique* B. Ainsi, l'ensemble $\mathcal{F}_{A,B}$ sera dit *famille de Moore implicative*.

Rappelons que l'ensemble $(\mathcal{P}(E), \cap, \cup)$ est un treillis. Comme $\mathcal{F}$ est un sous ensemble de $\mathcal{P}(E)$ stable par intersection donc $\mathcal{F}$ est un inf-demi-treillis. Par ailleurs, $\mathcal{F}$ contient un plus grand élément qui est l'ensemble E, donc la famille de Moore $\mathcal{F}$ est un treillis, par application du Théorème 2. Ainsi, nous avons le résultat suivant.

Théorème 3: [Mon03] Soit $\mathcal{F}$ une famille de Moore sur E. L'ensemble ordonné $(\mathcal{F}, \subseteq)$ est un treillis, avec les supremum et infimum définis par : pour tous $X, Y \in \mathcal{F}$, $X \wedge Y = X \cap Y$ et $X \vee Y = \cap\{F \in \mathcal{F} : X \cup Y \subseteq F\}$.

La réciproque du théoreme ci-dessus est vraie. En fait, nous avons le théorème suivant.

Théorème 4: [Mon03] Tout treillis est isomorphe à un treillis des fermés d'une famille de Moore.

Remarque 8: La notion duale de famille de Moore est définie de la façon suivante. Une famille d'ensembles $\mathcal{O}$ de parties E sera dite *duale de famille de Moore* ou encore un *système d'ouvertures*, si elle vérifie les deux conditions ci-dessous :

(O1) $\emptyset \in \mathcal{O}$;

(O2) $O_1, O_2 \in \mathcal{O}$ impliquent $O_1 \cup O_2 \in \mathcal{O}$.

Les éléments de $\mathcal{O}$ s'appellent les *ouverts* de $\mathcal{O}$.

2.3.2 Familles de Moore et Opérateurs de fermeture

Définition 13: Soit E un ensemble. Un *opérateur de fermeture* ou tout simplement *fermeture* sur E est une application ϕ définie sur $\mathcal{P}(E)$ satisfaisant aux trois conditions suivantes :

(F1) pour tous $A, B \subseteq E$, $A \subseteq B$ *implique* $\phi(A) \subseteq \phi(B)$ (isotonie) ;

(F2) pour tout $A \subseteq E$, $\phi\phi(A) = \phi(A)$ (idempotence) ;

(F3) pour tout $A \subseteq E$, $A \subseteq \phi(A)$ (extensivité).

Une *ouverture* sur E est une application ψ définie sur $\mathcal{P}(E)$ qui est à la fois isotone, idempotente et contractante (*i.e.*, pour tout $A \subseteq E$, $\psi(A) \subseteq A$).

Nous avons défini ci-dessus un opérateur de fermeture par trois axiomes (isotonie, idempotence et extensivité). Remarquons qu'il existe d'autres caractérisations des opérateurs de fermeture. Nous donnons ci-après trois de ces caractérisations.

Théorème 5: (Propriété d'indépendance de chemin [Plo73]) Soit ϕ une application extensive définie sur $\mathcal{P}(E)$. Alors, ϕ est une fermeture, si et seulement si, pour tous $A, B \subseteq E$, $\phi(A \cup B) = \phi(\phi(A) \cup \phi(B))$.

Théorème 6: (Relation de Morgado [Mor62]). Une application ϕ définie sur $\mathcal{P}(E)$ est une fermeture, si et seulement si, pour tous $X, Y \in \mathcal{P}(E)$, ϕ vérifie : $X \leq \phi(Y) \Leftrightarrow \phi(X) \leq \phi(Y)$.

Théorème 7: Une application ϕ définie sur $\mathcal{P}(E)$ est un opérateur de fermeture, si et seulement si, pour tous $X, Y \subseteq E, X \cup \phi(\phi(Y)) \subseteq \phi(X \cup Y)$ [Ise51].

La proposition ci-dessous établit la correspondance réciproque entre les notions d'opérateurs de fermetures et familles de More

Proposition 2: [GW99] Soit E un ensemble ordonné.

- Considérons une famille de Moore $\mathcal{F}$ sur E. Alors, l'application notée $\phi_{\mathcal{F}}$ définie par, pour tout $X \subseteq E$, $\phi_{\mathcal{F}}(X) = \cap\{F \in \mathcal{F} : X \subseteq F\}$ est un opérateur de fermeture.
- Réciproquement, soit ϕ un opérateur de fermeture sur E. Alors la famille d'ensembles $\mathcal{F}_\phi$ définie par $\mathcal{F}_\phi = \{F \subseteq E : \phi(F) = F\}$ (ensemble des points fixes de ϕ) est une famille de Moore sur E.
- On a les égalités suivantes $\phi_{\mathcal{F}_\phi} = \phi$ et $\mathcal{F}_{\phi_{\mathcal{F}}} = \mathcal{F}$.

Définition 14: Soient E un ensemble et $\mathcal{F}$ une famille de Moore finie sur E. Une partie Q de E est appelée ensemble *quasi-fermé* de $\mathcal{F}$, si $Q \notin \mathcal{F}$ et $\mathcal{F} \cup \{Q\}$ est une famille de Moore sur E. Etant donné $F \in \mathcal{F}$, Q est un ensemble F-quasi-fermé de $\mathcal{F}$ si Q est quasi-fermé et $\phi_{\mathcal{F}}(Q) = F$.

Exemple 5: Soit $E = \{a, b, c, d, e\}$. Nous avons vu, dans l'exemple 3, que la famille $\mathcal{F} = \{\emptyset, a, b, d, de, bcd, abcde\}$ est une famille de Moore sur l'ensemble E. Alors "ab" est quasi-femé et $\phi_{\mathcal{F}}(ab) = abcde$. Par contre "$ae$" n'est pas quasi-fermé, car $de \cap ae = e \notin \mathcal{F} \cup ae$.

Remarque 9: Pour une famille de Moore $\mathcal{F}$ sur un ensemble E, tout élément A minimal de $(\mathcal{P}(E) \setminus \mathcal{F})$ est quasi-fermé (en effet, $A \in min(\mathcal{P}(E) \setminus \mathcal{F})$ implique $(F \cap A) \in \mathcal{F}$ pour tout $F \in \mathcal{F}$ tel que $A \not\subseteq F$).

Proposition 3: [CM03] Une partie Q de E est quasi-fermé de $\mathcal{F}$, si et seulement si Q n'est pas fermé et pour tout $X \subset Q, \phi_{\mathcal{F}}(X) \subset \phi_{\mathcal{F}}(Q)$ implique $\phi_{\mathcal{F}}(X) \subset Q$.

Récemment Diatta [Dia05] a proposé une autre caractérisation des ensembles quasi-fermés d'une famille de Moore finie selon la proposition ci-dessous.

Proposition 4: [Dia05] Une partie Q de E est quasi-fermé d'une famille de Moore $\mathcal{F}$, si et seulement si $Q \notin \mathcal{F}$ et pour tout $X \subset Q$, ou bien $\phi_{\mathcal{F}}(X) \subset Q$ ou bien $Q \subset \phi_{\mathcal{F}}(X)$.

Définition 15: Soit $\mathcal{F}$ une famille de Moore sur E. Un ensemble $C \subseteq E$ sera appelé un ensemble *critique*, si C est un minimal $\phi_{\mathcal{F}}(C)$-quasi-fermé [Day92].

Les ensembles critiques triviaux sont spécifiés par la proposition ci-dessous.

Proposition 5: Soit $\mathcal{F}$ une famille de Moore sur un ensemble E. Alors tout élément minimal de $\mathcal{P}(E) \setminus \mathcal{F}$ est un ensemble critique de $\mathcal{F}$.

Exemple 6: Considérons la famille de Moore $\mathcal{F} = \{\emptyset, a, b, e, ae, cd, bcd, abcde\}$ de l'exemple 3. Alors "c" est un ensemble critique de $\mathcal{F}$. Par contre "bc" est un ensemble quasi-fermé de $\mathcal{F}$, mais n'est pas critique car $\phi_{\mathcal{F}}(c) = \phi_{\mathcal{F}}(bc) = bcd$.

Nous rappelons ci-dessous la caractérisation dite récursive des ensembles critiques d'une famille de Moore.

Proposition 6: [Cas99, CM03] Soit $\mathcal{F}$ une famille de Moore sur un ensemble E. $C \subset E$ est un ensemble critique, si et seulement si pour tout $C' \subset C$, C' critique alors $\phi_{\mathcal{F}}(C') \subset C$.

Une autre caractérisation des ensembles critiques d'une famille de Moore est donnée par la proposition suivante.

Proposition 7: [Dia05] Soit $\mathcal{F}$ une famille de Moore sur un ensemble E. Une partie C de E est critique, si et seulement si $C \notin \mathcal{F}$ et pour tout $C' \subset C$, ou bien $\phi_{\mathcal{F}}(C') \subset C$ ou bien $C \subset \phi_{\mathcal{F}}(C')$ et il existe $C'' \subset C'$ tel que $\phi_{\mathcal{F}}(C'')$ intersecte proprement C' ($\phi_{\mathcal{F}}(C'') \cap C' \neq \emptyset$, $\phi_{\mathcal{F}}(C'') \setminus C' \neq \emptyset$ et $C' \setminus \phi_{\mathcal{F}}(C'') \neq \emptyset$).

Définition 16: Soit $\mathcal{F}$ une famille de Moore sur un ensemble E. Notons $\mathcal{C}$ l'ensemble de tous les ensembles critiques de $\mathcal{F}$. L'ensemble $\mathcal{B}_{\mathcal{C}} = \{\mathcal{F}_{C,\phi(C)}, C \in \mathcal{C}\}(\mathcal{F}_{C,\phi(C)}$ (famille de Moore implicative) est appelé la base canonique de $\mathcal{F}$.

2.3.3 Familles de Moore et Systèmes implicatifs

Nous avons evoqué au début de cette section que la notion de famille de Moore est équivalente à la notion de systèmes implicatifs. Ces derniers jouent un rôle fondamental dans plusieurs domaines comme l'analyse de données, la base de données relationnelle et l'extraction de connaissances. Le problème de ces différents domaines est de trouver le système minimum d'implications permettant de générer toutes les implications entre les entités dans un contexte de données. Le présent paragraphe concerne les correspondances entre les notions de familles de Moore et systèmes implicatifs.

Définition 17: Soit E un ensemble fini. Un *système implicatif* sur E, noté Σ, est une relation binaire définie sur $\mathcal{P}(E)$: $\Sigma \subseteq \mathcal{P}(E) \times \mathcal{P}(E)$. Si $(X, Y) \in \Sigma$, on écrit $X \rightarrow_{\Sigma} Y$ ou tout simplement $X \rightarrow Y$ et on lit X *implique* Y ou $X \rightarrow Y$ est une implication de Σ.

Un système implicatif Σ sur E est dit *complet* s'il vérifie, pour tous $X, Y, Z, T \subseteq E$, les trois conditions suivantes :

(I1) $X \supseteq Y$ implique $X \rightarrow Y$;

(I2) $X \rightarrow Y$ et $Y \rightarrow Z$ impliquent $X \rightarrow Z$ (transitivité) ;

(I3) $X \rightarrow Y$ et $Z \rightarrow T$ impliquent $X \cup Z \rightarrow Y \cup T$ (augmentation).

Définition 18: Soit Σ un système implicatif défini sur un ensemble E.

(1) Une implication $X \rightarrow_{\Sigma} Y$ est dite *propre* (resp. *triviale*) si $X \cap Y = \emptyset$ (resp. $Y \subseteq X$).

(2) Soit $y \in E$. Toute implication de la forme $X \rightarrow_{\Sigma} \{y\}$ s'appelle *implication élémentaire*.

Remarque 10: Toute implication $X \rightarrow_{\Sigma} Y$ est équivalente à la conjonction d'implications $X \rightarrow_{\Sigma} \{y\}$, pour tout $y \in Y$, grâce aux propriétés (I1, I2, I3).

La proposition ci-dessous montre les correspondances réciproques entre les notions de familles de Moore et systèmes implicatifs complets.

Proposition 8: [CM03]

(i) Soit $\mathcal{F}$ une famille de Moore sur un ensemble E. $\Sigma_{\mathcal{F}}$ défini par $\Sigma_{\mathcal{F}} = \{X \rightarrow Y : \mathcal{F}^X \subseteq \mathcal{F}^Y\}$ avec $\mathcal{F}^X = \{F \in \mathcal{F} : X \subseteq F\}$ est un système implicatif complet.

(ii) Réciproquement, Soit Σ un système implicatif complet sur E. La famille $\mathcal{F}_\Sigma$ définie par $\mathcal{F}_\Sigma = \{F \subseteq E : X \subseteq F \text{ et } X \rightarrow_\Sigma Y \text{ impliquent } Y \subseteq F\}$ est une famille de Moore.

Nous avons déjà vu les correspondances entre les notions de familles de Moore et opérateurs de fermeture, les notions de familles de Moore et systèmes implicatifs complets. Donc, il est logique que les deux notions opérateurs de fermeture et systèmes implicatifs sont équivalentes. La proposition ci-dessous établit les relations directes entre ces deux notions.

Proposition 9: [CM03]

(i) Soit ϕ un opérateur de fermeture défini sur E. Σ_ϕ définie par $\Sigma_\phi = \{X \rightarrow Y : Y \subseteq \phi(X)\}$ est un système implicatif complet.

(ii) Réciproquement, considérons Σ un système implicatif complet. L'application ϕ_Σ définie par, pour tout $X \subseteq E$, $\phi_\mathcal{F}(X) = \bigcup\{x \in E, X \rightarrow_\Sigma \{x\}\}$ est un opérateur de fermeture.

Définition 19: Soient Σ un système implicatif complet et $\mathcal{F}_\Sigma$ la famille de Moore associée à Σ. Une implication $X \rightarrow_\Sigma Y$ est dite Σ-*critique*, si X est un ensemble critique pour la famille de Moore $\mathcal{F}_\Sigma$ et $Y = \phi_\mathcal{F}(X) - X$ (avec ϕ_Σ est un opérateur associé à $\mathcal{F}_\Sigma$)

Théorème 8: [GD86] Soit Σ un système implicatif complet sur un ensemble E. L'ensemble de toutes les implications Σ-critiques est la base minimale pour Σ. Par ailleurs toute base peut être obtenue à partir de cette base minimale.

2.4 Correspondances de Galois

Cette section concerne les correspondances de Galois. Elle joue un rôle important dans la théorie des ensembles ordonnés. Nous en présentons d'abord quelques généralités avant de voir le cas particulier de correspondance de Galois associée à une relation binaire.

2.4.1 Généralités

Définition 20: Soient $(E, \leq)$, $(F, \leq)$ deux ensembles ordonnés et $f : E \to F$ et $g : F \to E$ deux applications. Le couple d'applications (f, g) sera dit *correspondance de Galois* entre E et F si, pour tous $x, x' \in E, y, y' \in F$, les trois conditions suivantes sont vérifiées :

(G1) $x \leq x'$ implique $f(x) \geq f(x')$ (antitonie) ;

(G2) $y \leq y'$ implique $g(y) \geq g(y')$ (antitonie) ;

(G3) $x \leq g \circ f(x)$ *et* $y \leq f \circ g(y)$ (extensivité).

L'application f (resp. g) est appelée *application galoisienne* de E vers F (resp. de F vers E).

Proposition 10: Soient $(E, \leq)$, $(F, \leq)$ deux ensembles ordonnés et $f : E \to F$ et $g : F \to E$ deux applications. Le couple (f, g) est une correspondance de Galois, si et seulement si :

(G4) pour tous $x \in E$ et $y \in F$, $x \leq g(y) \Leftrightarrow y \leq f(x)$.

Proposition 11: Soit (f, g) une correspondance de Galois. On a les égalités suivantes :

$$f = f \circ g \circ f \text{ et } g = g \circ f \circ g.$$

Ainsi, une correspondance de Galois (f, g) sera dit couple *involutif*.

Proposition 12: Soit (f, g) une correspondance de Galois entre deux ensembles E et F. Notons $\phi = g \circ f$ et $\phi' = f \circ g$. Les applications ϕ et ϕ' sont des opérateurs de fermetures respectivement sur E et F.

Définition 21: Soient E et F deux ensembles ordonnés. Considérons les applications $f : E \to F$ et $g : F \to E$. L'application f (resp. g) sera dite application *résiduée* (resp. *résiduelle*) si pour tous $x, x' \in E$ et $y, y' \in F$, les trois conditions suivantes sont vérifées :

(R1) $x \leq x'$ implique $f(x) \leq f(x')$ (isotonie) ;

(R2) $y \leq y'$ implique $g(y) \leq g(y')$ (isotonie) ;

(R3) $x \leq g \circ f(x)$ (extensivité) et $y \geq f \circ g(y)$ (contraction).

Proposition 13: Une application f de E vers F (resp. g de F vers E) est résiduée (resp. résiduelle) si et seulement si f (resp. g) est une application galoisienne de E dans F^d (resp. F^d vers E) où F^d est le dual de F.

Donc, la notion d'applications résiduée/ résiduelle et celle de correspondance de Galois sont équivalentes.

2.4.2 Correspondances de Galois associées à une relation binaire

Soient E et F deux ensembles finis et $\mathcal{R}$ une relation binaire de E vers F. Définissons deux fonctions $f_{\mathcal{R}}$ et $g_{\mathcal{R}}$ de la façon suivante :

$f_{\mathcal{R}} : \mathcal{P}(E) \to \mathcal{P}(F)$
$X \longmapsto f_{\mathcal{R}}(X) = \bigcap_{x \in X} \{y \in F : x\mathcal{R}y\} = \{y \in F : \text{pour tout } x \in X, x\mathcal{R}y\}$

$g_{\mathcal{R}} : \mathcal{P}(F) \to \mathcal{P}(E)$
$Y \longmapsto g_{\mathcal{R}}(Y) = \bigcap_{y \in Y} \{x \in E : x\mathcal{R}y\} = \{x \in E : \text{pour tout } y \in Y, x\mathcal{R}y\}$

Théorème 9: [GW99]

- Le couple $(f_{\mathcal{R}}, g_{\mathcal{R}})$ est une correspondance de Galois.
- Réciproquement, si (f, g) est une correspondance de Galois entre les ensembles $\mathcal{P}(E)$ et $\mathcal{P}(F)$ alors $\mathcal{R}_{(f,g)} = \{(x, y) \in E \times F : x \in g(\{y\})\} = \{(x, y) \in E \times F : y \in f(\{x\})\}$ est une relation binaire de E vers F.
- Par ailleurs, nous avons les égalités suivantes $f_{\mathcal{R}_{(f,g)}} = f$, $g_{\mathcal{R}_{(f,g)}} = g$ et $\mathcal{R}_{(f_{\mathcal{R}}, g_{\mathcal{R}})} = \mathcal{R}$.

Le corollaire suivant est une conséquence de la Proposition 12.

Corollaire 1: Les applications $\phi = g_{\mathcal{R}} \circ f_{\mathcal{R}}$ et $\phi' = f_{\mathcal{R}} \circ g_{\mathcal{R}}$ sont des opérateurs de fermetures respectivement sur $\mathcal{P}(E)$ et $\mathcal{P}(F)$.

Ce type de correspondance de Galois joue un rôle important en analyse formelle de concepts (A.F.C.). Cette notion d'A.F.C. fournit un cadre théorique fondamental pour la fouille des règles d'association d'un contexte binaire. En A.F.C., l'ensemble E désigne un ensemble fini d'entités et F un ensemble fini d'attributs ou variables. En fait, nous avons la définition suivante.

Définition 22:
- Un *contexte formel* est un triplet $\mathbb{K} = (E, F, \mathcal{R})$ où E est ensemble fini d'entités, F un ensemble fini d'attributs et $\mathcal{R}$ une relation binaire de E vers F.
- Soit $(X, Y) \in \mathcal{R}$. Le couple (X, Y) sera dit *concept formel* si X est un fermé de ϕ (*i.e.* $\phi(X) = X$) et $Y = f_{\mathcal{R}}(X)$ avec $\phi = g_{\mathcal{R}} \circ f_{\mathcal{R}}$.

Exemple 7: Le Tableau 2.1 présente un contexte formel avec $E = \{X_1, X_2, X_3, X_4, X_5, X_6\}$ et $F = \{Y_1, Y_2, Y_3, Y_3, Y_4, Y_5\}$. Dans ce tableau,

$E \setminus F$	Y_1	Y_2	Y_3	Y_4	Y_5
X_1	×	×	×	×	
X_2		×	×	×	
X_3	×		×	×	×
X_4	×	×	×		×
X_5		×		×	×
X_6	×			×	×

Tab. 2.1: *Contexte formel*

on écrit $(X, Y) \in \mathcal{R}$ lorsqu'une case contient $\times$, par exemple (X_1, Y_3), $(X_6, Y_1) \in \mathcal{R}$ et $(X_5, Y_3) \notin \mathcal{R}$.

Désignons par $\mathcal{C}_F$ l'ensemble de tous les concepts d'un contexte formel $\mathbb{K}$. Définissons une relation d'ordre $\leq$ sur $\mathcal{C}_F$, pour tous $(X_1, Y_1), (X_2, Y_2) \in \mathcal{C}_F$, par $(X_1, Y_1) \leq (X_2, Y_2)$, si et seulement si $X_1 \subseteq X_2$ (ou, équivalente $Y_2 \subseteq Y_1$). Nous avons la proposition suivante.

Proposition 14: L'ensemble $\mathcal{C}_F$ muni de la relation d'ordre $\leq$ forme un treillis complet appelé *treillis de Galois*. Et on a $(X_i, Y_j) \subseteq \mathcal{C}_F$, $\bigwedge(X_i, Y_j) = (\bigcap X_i, \phi'(\bigcup Y_j))$ et $\bigvee(X_i, Y_j) = (\phi(\bigcup X_i), \bigcap Y_j))$

2.5 Conclusion

Nous avons présenté dans ce chapitre quelques notions mathématiques qui seront utiles dans la suite : treillis, familles de Moore et correspondances de Galois. L'utilisation de ces trois notions permet de pallier certains problèmes de la fouille des règles d'association. En particulier, elles permettent de construire des représentations condensées de règles d'association extraites à partir d'un contexte formel.

3. RÈGLES D'ASSOCIATION

3.1 Introduction

L'Extraction de Connaissances à partir (des bases) de Données (E.C.D. ou E.C.B.D.) désigne le processus non trivial d'extraction des informations implicites, précédemment inconnues et potentiellement utiles concernant les données stockées dans les bases de données [PS91] ou encore le processus non trivial d'identification de structures inconnues, valides et potentiellement exploitables dans les bases de données [FPSS96]. On distingue cinq principaux problèmes en E.C.D. : l'extraction des règles d'association [AIS93], la classification [EP96], le clustering ou la classification non supervisée [CS96], les séries chronologiques [WMSZ94] et la généralisation des données [SA96].

L'extraction des règles d'association est devenue aujourd'hui l'une des tâches les plus populaires de la fouille de données, et ce, depuis les travaux de Agrawal et al. [AIS93, AS94]. L'analyse de panier de ménagère est l'une des applications typiques de l'extraction des règles d'association. Elle a pour but de dégager les relations intelligibles entre les attributs dans une base de données. Dans le cas de l'analyse du panier de ménagère, l'extraction des règles d'association permet d'analyser les tickets de caisse des clients particuliers afin de comprendre leurs habitudes de consommation, agencer les rayons du magasin, organiser les promotions, gérer les stocks etc. dans le naturel but d'améliorer le profit. Dans la base de données de vente, une *transaction* consiste à un ensemble d'articles achetés par un client particulier, appelés *items* ou *attributs*. Ainsi, une base de données est un ensemble de transactions qu'on appelle aussi base *transactionnelle*. Dans un tel contexte, une règle d'association est une implication conditionnelle entre des ensembles d'attributs dans une base transactionnelle. En d'autres termes, étant donné un ensemble d'attributs, le but de l'extraction des règles d'association est de découvrir "si l'occurrence d'un tel ensemble dans une transaction est associée à l'occurrence d'un autre ensemble d'attributs". Par exemple, "70% des clients achetant du lait et du thé achétent aussi du pain" est une règle

d'association associant les attributs lait et thé à l'attribut pain. Ce problème est loin d'être trivial, vu le nombre exponentiel du nombre d'attributs de la base transactionnelle. Par ailleurs, une base transactionnelle stocke des millions de transactions sur des milliers d'attributs.

Ce chapitre concerne les processus d'extraction des règles d'association d'un contexte de la fouille de données binaires. Il est organisé de la façon suivante. Après avoir exposé la problématique de fouille des règles d'association dans la Section 3.2, nous donnons quelques définitions et présentons les principes de base de la fouille des règles d'association dans la Section 3.3. La Section 3.4 concerne l'extraction des représentations condensées des motifs fréquents. Nous concluons par une courte conclusion dans la Section 3.5.

3.2 Problématiques

Le cadre classique de problème de la fouille des règles d'association introduite dans [AIS93, AS94] peut être décrit de la façon suivante. Soient $\mathcal{A} = \{x_1, x_2, ..., x_m\}$ un ensemble de m articles ou attributs dits aussi variables, et $\mathcal{E} = \{e_1, e_2, ..., e_n\}$ un ensemble de n transactions ou entités définies sur l'ensemble d'attributs. Un sous-ensemble X de $\mathcal{A}$ est appelé motif. Chaque entité e_i consiste à un ensemble d'attributs de $\mathcal{A}$. A chaque entité, on associe un identificateur unique appelé TID (Transaction IDentifier). Une base de transactions $\mathbb{K}$ est donc un ensemble de couples formés d'un identificateur de transactions et de la transaction proprement dite.

Définition 23:

- Le Support d'un motif est le rapport du nombre de transactions contenant ce motif par le nombre de transaction dans la base transactionnelle.

- Une règle d'association est une implication de la forme $X \rightarrow Y$ (où $Y \neq \emptyset$) exprimant le fait que les attributs dans X tendent à apparaître avec ceux dans Y. Dans ce cas, X est appelé la *prémisse* de la règle et Y son *conséquent*.

- Le *Support* d'une règle d'association $X \rightarrow Y$ est égal au Support de motif $X \cup Y$. La *Confiance* d'une règle d'association est la proportion de transactions contenant le conséquent parmi celles contenant la prémisse de la règle.

- La qualité d'une règle d'association est classiquement évaluée par le couple de mesures Support et Confiance. Pour un seuil minimum de Support "minsupp" et un seuil minimum de Confiance "minconf", une règle est dite valide si son Support et sa Confiance dépassent respectivement les seuils de Support et de Confiance préalablement fixés.

Le problème de la fouille des règles d'association d'une grande base de données est alors traditionnellement défini de la façon suivante. Connaissant les seuils minimum de Support minsupp et de Confiance minconf, trouver toutes les règles d'association valides de cette base de données. Grossièrement, ce problème est subdivisé en deux sous-probèmes [HGN00] :

1. trouver tous les motifs fréquents, *i.e.* les motifs dont le Support est supérieur ou égal au minsupp, seuil fixé par l'utilisateur ;

2. générer toutes les règles d'association valides dérivant des motifs fréquents.

La solution du second sous-problème étant directe, plus d'éfforts de recherche ont été consacrés au premier sous-problème. Il existe plusieurs algorithmes de résolution de problème de la fouille des règles d'association dans la littérature, entre autres, APRIORI, APRIORITID [AS94], Max-Miner [Bay98], etc.

Notons toutefois que l'espace de recherche des motifs fréquents est clairement très large. En particulier si les données sont fortement corrélées ou si le seuil minimum de Support est très petit, l'extraction des motifs fréquents est presque impossible. Plusieurs solutions [Zak00b, ZPOL97, HPY00] ont été proposées pour améliorer les algorithmes de génération de ces motifs. Dans ces travaux, le volume de la base de données et le nombre abondant des motifs fréquents sont considérés comme les aspects les plus coûteux de la fouille, et donc plus d'efforts ont été consacrés pour minimiser le nombre de parcours à travers la base de données.

Pour pallier le problème de l'extraction des motifs fréquents, l'idée de trouver une *repésentation condensée* de ces motifs préoccupe plusieurs chercheurs travaillant dans le domaine de la fouille des règles d'association. Plusieurs propositions ont été soumises dans cette direction [PBTL99a, PHM00, BB00, BBR00, KG00, BBR03].

Une représentation condensée est une collection des motifs fréquents non redondants à partir de laquelle, on peut dériver l'ensemble de tous les motifs fréquents et leurs Supports. Dans les cas pratiques, le nombre d'éléments

d'une représentation condensée est significativement plus petit que celui de l'ensemble de tous les motifs fréquents. Par ailleurs, l'extraction d'une représentation condensée est toujours possible même dans le cas des données fortement corrélées ou le cas le seuil minimum de Support a une valeur très faible.

3.3 Définitions et principes de base

3.3.1 Définitions

Soient $\mathcal{A} = \{x_1, x_2, ..., x_m\}$ un ensemble de m articles ou attributs dits aussi variables et $\mathcal{E} = \{e_1, e_2, ..., e_n\}$ un ensemble de n transactions ou entités définies sur l'ensemble d'attributs, *i.e.*, une entité e_i consiste à un ensemble d'attributs de $\mathcal{A}$. Un sous ensemble X de $\mathcal{A}$ est appelé motif.

Définition 24:

- $X \subseteq \mathcal{A}$ sera appelé *k-motif* si la taille de X est égale à k.
- Soient $E \in \mathcal{E}$ une entité et X un motif. On dit que E contient X si $X \subseteq E$.
- On définit le *Support* d'un motif X par $\text{Supp}(X) = \frac{|\{E \in \mathcal{E} | X \subseteq E\}|}{|\{E \in \mathcal{E}\}|}$, où $|A|$ désigne la cardinalité d'un ensemble A. Le Support d'un motif est donc le rapport de la cardinalité de l'ensemble des transactions qui contiennent tous les attributs de X par la cardinalité de l'ensemble de toutes les transactions.
- Un motif X sera dit *fréquent* si $\text{Supp}(X) \geq \text{minsupp}$, où minsupp est un seuil minimum de Support, fixé par l'expert.
- Une *règle d'association* est une implication conditionnelle $X_1 \rightarrow X_2$, où X_1 et X_2 sont deux motifs tels que $X_2 \neq \emptyset$; X_1 sera dit la *prémisse* de la règle et X_2 son *conséquent*.
- On définit le Support d'une règle d'association $X_1 \rightarrow X_2$ par $\text{Supp}(X_1 \rightarrow X_2) = \text{Supp}(X_1 \cup X_2)$. Il indique la proportion d'entités contenant à la fois la prémisse et le conséquent de la règle.
- La *Confiance* d'une règle d'association $X_1 \rightarrow X_2$, notée $\text{Conf}(X_1 \rightarrow X_2)$, est définie par $\text{Conf}(X_1 \rightarrow X_2) = \frac{\text{Supp}(X_1 \rightarrow X_2)}{\text{Supp}(X_1)}$. Elle indique la proportion d'entités contenant le conséquent parmi celles qui contiennent la prémisse.

- Une règle d'association $X_1 \rightarrow X_2$ est dite *valide* au sens des mesures de qualité $(\text{Supp}, \text{Conf})$ si $X_1 \cup X_2$ est fréquent et si $\text{Conf}(X_1 \rightarrow X_2) \geq \text{minconf}$, où minconf est un seuil minimum de Confiance, fixé par l'utilisateur.

3.3.2 Motifs fréquents : algorithme APRIORI

Agrawal et al. ont proposé dans [AS94], le premier algorithme, appelé APRIORI, pour l'extraction des règles d'association dans les bases de données transactionnelles.

La propriété d'anti-monotonicité de Support est utilisée dans l'algorithme APRIORI pour élaguer les motifs non fréquents d'une base de données volumineuse. Nous rappelons ci-dessous la propriété d'antimonotonicité.

Définition 25: Une propriété ρ est anti-monotone, si et seulement si pour tous motifs X et Y tels que, $\rho(X)$ et $Y \subseteq X$ impliquent $\rho(Y)$.

Dans le cas de Support, nous avons le théorème suivant.

Théorème 10: Soient X et Y deux motifs. On a : si $X \subseteq Y$, alors $\text{Supp}(X) \geq \text{Supp}(Y)$.

Soit minsupp $\in [0, 1]$ un seuil minimum de Support, fixé par l'utilisateur. Rappelons qu'un motif X est dit fréquent si $\text{Supp}(X) \geq \text{minsupp}$. Une conséquence du Théorème 10 est la suivante.

Corollaire 2: Soit $X \subseteq \mathcal{A}$ un motif.

(i) Si X est fréquent, alors pour tout motif X_1, tel que $X_1 \subseteq X$, X_1 est aussi fréquent, *i.e.*, tout sous-motif d'un motif fréquent est un motif fréquent

(ii) Si X est non fréquent, alors pour tout motif X_2, tel que $X \subseteq X_2$, X_2 est aussi non fréquent, *i.e.*, tout sur-ensemble d'un motif non fréquent est non fréquent.

Étant donné une base de données transactionnelles $\mathcal{E}$, le problème consiste à trouver comment générer toutes les règles d'association valides liant les motifs fréquents entre eux.

L'algorithme APRIORI se base essentiellement sur la propriété d'antimonotonicité existant entre les motifs : si un motif est non fréquent, alors tous

ses sur-ensembles ne sont plus testés. Les motifs candidats à être fréquents sont les motifs dont tous les sous-ensembles sont connus fréquents. Cette propriété est utilisée à chaque itération afin de déterminer les motifs candidats à considérer. Pour optimiser la génération des motifs candidats et le calcul de leurs Supports, les motifs sont ordonnés par ordre lexicographique. [1]

Rappelons que l'ensemble $\mathcal{P}(\mathcal{A})$ de tous les motifs X d'une base de données forme un treillis complet. L'algorithme APRIORI utilise une approche itérative par niveau pour générer tous les motifs fréquents. Pour cela, les treillis des motifs est exploré en largeur d'abord. APRIORI effectue, à chaque itération k, un passage dans la base de transactions afin de calculer le Support de chaque k-motif. Dans la suite, l'ensemble des k-motifs candidats (*i.e.*, dont on ne connaît pas encore le Support dans la base de données) sera noté par $\mathcal{C}_k$ et l'ensemble des k-motifs fréquents de taille k par $\mathcal{F}_k$.

Le pseudo-code pour générer tous les motifs fréquents est présenté dans l'algorithme 1 et les notations utilisées sont données dans le Tableau 3.1.

$\mathcal{C}_k$	Ensemble des k-motifs candidats et leurs Supports.
$\mathcal{F}_k$	Ensemble de k-motifs fréquents et leurs Supports

Tab. 3.1: *Notations utilisées dans l'Algorithme 1*

Algorithme 1 (APRIORI):
Entrée: *Base de transactions* $\mathbb{K}$, minsupp.
Sortie: $\mathcal{F}$ *: ensemble de tous les motifs fréquents et leurs Supports.*

1: $\mathcal{F}_1 \leftarrow \{1\text{-}motifs\ fréquents\}$
2: ***for*** $(k \leftarrow 2; \mathcal{F}_{k-1} \neq \emptyset; k++)$ ***do***
3: $\quad \mathcal{C}_k \leftarrow Apriori\text{-}Gen(\mathcal{F}_{k-1})$
4: $\quad$ ***for all*** $E \in \mathcal{E}$ ***do***
5: $\quad\quad \mathcal{C}_E \leftarrow sous\text{-}ensemble(\mathcal{C}_k, E) // \ \mathcal{C}_E = \{c \in \mathcal{C}_k, c \subseteq E\}$
6: $\quad\quad$ ***for all*** $c \in \mathcal{C}_E$ ***do***
7: $\quad\quad\quad \text{Supp}(c)++$
8: $\quad\quad$ ***end for***

[1] Un ordre lexicographique est une relation d'ordre sur E^k, où E est un ensemble totalement ordonné et k un entier. On la définit de la façon suivante : $(x_1, x_2, ..., x_k) \leq (y_1, y_2, ..., y_k)$, si et seulement si il existe i tel que pour tout $j < i$, $x_j = y_j$ et $x_i \leq y_i$.

9: $\mathcal{F}_k \leftarrow \{c \in \mathcal{C}_E, \text{Supp}(c) \geq \text{minsupp}\}$
10: ***end for***
11: *Retourner* $\mathcal{F} \leftarrow \bigcup_k \mathcal{F}_k$
12: ***end for***

Les motifs fréquents sont calculés de façon itérative, dans l'ordre ascendant suivant leurs tailles. Cet algorithme prend l itérations, l étant la taille maximale de motifs fréquents. Pour chaque itération $k \leq l$, la base de données est parcourue une fois et tous les motifs fréquents de taille k sont calculés. La ligne 1 consiste à trouver tous les 1-motifs fréquents. L'algorithme alterne ensuite la génération des candidats et calcule les fréquents parmi ces candidats (lignes 2 à 10). Pour cela, à l'itération k, l'ensemble $\mathcal{F}_{k-1}$ des $(k-1)$-motifs fréquents correspondant aux motifs de niveau $(k-1)$ du treillis (utilisé à l'étape précédente), est utilisé pour générer l'ensemble de $\mathcal{C}_k$ des k-motifs candidats.

Algorithme 2 (Apriori-Gen):
Entrée : $\mathcal{F}_{k-1}$
Sortie : $\mathcal{C}_k$
1: $\mathcal{C}_k \leftarrow \emptyset$
2: ***for all*** $p \in \mathcal{F}_{k-1}$ ***do***
3: ***for all*** $q \in \mathcal{F}_{k-1}$ ***do***
4: ***if*** $p(1) = q(1), p(2) = p(2), ..., p(k-2) = q(k-2), p(k-1) < q(k-1)$ ***then***
5: $c \leftarrow p \cup q(k-1)$
6: ***end if***
7: ***for all*** $s \subseteq c$ (*avec* s *un* $(k-1)$*-motif*) ***do***
8: ***if*** $s \in \mathcal{F}_{k-1}$ ***then***
9: $\mathcal{C}_k \leftarrow \mathcal{C}_k \cup \{c\}$
10: ***end if***
11: ***end for***
12: ***end for***
13: *Retourner* $\mathcal{C}_k$
14: ***end for***

La procédure Apriori-Gen appelée en ligne 3 prend $\mathcal{F}_{k-1}$ comme donnée d'entrée et $\mathcal{C}_k$ comme résultat. L'initialisation de $\mathcal{C}_k$ à l'ensemble vide est faite en ligne 1. Ensuite, une jointure est effectuée entre les éléments de

$\mathcal{F}_{k-1}$ (lignes 2 à 6). Deux motifs p et q de $\mathcal{F}_{k-1}$ forment un motif c si, et seulement s'ils ont $(k-2)$ attributs (dans le prefixe)en commun, ce qui est exprimé en utilisant l'ordre lexicographique dans la condition de la ligne 4 de l'algorithme Apriori-Gen. Les étapes suivantes (lignes 7 à 11) assurent, après avoir généré un candidat de taille k à partir de deux $(k-1)$-motifs fréquents, que tous les sous-ensembles du nouveau candidat sont fréquents.

Une fois l'ensemble $\mathcal{C}_k$ des motifs candidats est calculé, la base de transactions est parcourue afin de trouver le Support de chaque candidat. Les étapes (lignes 4 à 10) recherchent parmi les candidats de $\mathcal{C}_k$ ceux qui sont contenus dans la transaction E. Si c'est le cas, alors le Support de ces candidats est augmenté (ligne 7). Parmi les candidats, seuls ceux qui ont le Support supérieur à minsupp sont retenus.

Remarque 11: On retrouve dans la littérature un large éventail d'algorithmes considérés comme variantes de l'APRIORI, permettant de générer tous les motifs fréquents d'une base transactionnelle, entre autres, APRIORI-TID [AS94], PARTITION [SON95], DIC [BMUT97], SAMPLING [Toi94], ECLAT [ZPOL97], FP-growth [HPY00].

3.3.3 Génération des règles d'association

La génération des règles est beaucoup moins coûteuse que la génération des motifs fréquents, car il n'est plus nécessaire de faire le parcours coûteux de la base des données. Pour générer les règles d'association, on considère l'ensemble $\mathcal{F}$ des motifs fréquents trouvés dans la phase précédente. Pour chaque motif fréquent I, on considère tous ses sous-ensembles (tous fréquents d'après la propriété d'antimonotonicité). À partir de ces sous-ensembles fréquents, on génère toutes les règles $I_1 \rightarrow I \setminus I_1 (I_1 \subset I)$ telles que leurs Confiances respectives dépassent le seuil minimum de Confiance. Agrawal et Srikant ont proposé dans [AS94] une optimisation de la génération des règles d'association. Cette optimisation est basée sur la proposition suivante. Elle permet de ne pas considérer tous les ensembles possibles des motifs fréquents.

Proposition 15: Soit I un motif fréquent. Nous avons :

$$\forall I_1 \subset I, I_1 \neq \emptyset [I_1 \rightarrow I \setminus I_1] \text{est valide} \Rightarrow \forall I_2 \subset I_1, I_2 \neq \emptyset, [I_2 \rightarrow I \setminus I_2] \text{ est valide}$$

L'algorithme de génération des règles d'association valides, proposé dans [AS94] se base sur le résultat de la proposition ci-dessus. Il fonctionne de la

façon suivante. Étant donné un motif I, il génère toutes les règles ayant 1-motif comme conséquent. Les conséquents de ces règles sont ensuite combinés en réutilisant la fonction Apriori-Gen, et ce pour générer les conséquents possibles à 2-motifs pouvant apparaître dans une règle générée à partir de I et ainsi de suite. L'algorithme 3 génère les règles utilisant l'idée présentée ci-dessus. $\mathcal{F}$ présente l'ensemble des motifs fréquents et H_m celui de m-motifs conséquents de règles.

Algorithme 3 (Apriori-Gen-R.A.):
Entrée : $\mathcal{F}, \text{minconf}$
Sortie : $\mathcal{R}$ *ensemble des règles d'association*

1: $\mathcal{R} \leftarrow \emptyset$
2: ***for all*** k*-motif* $l_k \in \mathcal{F}, k \geq 2$ ***do***
3: $H_1 \leftarrow \{1 - \textit{motifs fréquents sous-ensembles de } l_k\}$
4: ***for all*** $h_1 \in H_1$ ***do***
5: $\text{Conf} \leftarrow \frac{\text{Supp}(l_k)}{\text{Supp}(l_k - h_1)}$
6: ***if*** $\text{Conf} \geq \text{minconf}$ ***then***
7: $\mathcal{R} \leftarrow \mathcal{R} \cup \{r : l_k - h_1 \rightarrow h_1\}$
8: ***else***
9: $H_1 \leftarrow H_1 - \{h_1\}$
10: ***end if***
11: ***end for***
12: $Gen - rules(l_k, H_1)$
13: $Retourner\ \mathcal{R}$
14: ***end for***

Algorithme 4 (Gen-Rules):

1: ***if*** $k > m + 1$ ***then***
2: $H_m \leftarrow Apriori - Gen(H_m)$
3: ***for all*** $h_{m+1} \in H_{m+1}$ ***do***
4: $\text{Conf} \leftarrow \frac{\text{Supp}(l_k)}{\text{Supp}(l_k - h_{m+1})}$
5: ***if*** $\text{Conf} \geq \text{minconf}$ ***then***
6: $\mathcal{R} \leftarrow \mathcal{R} \cup \{r : l_k - h_{m+1} \rightarrow h_{m+1}\}$
7: ***else***
8: $H_{m+1} \leftarrow H_{m+1} - \{h_{m+1}\}$
9: ***end if***
10: ***end for***

11: $Gen - rules(l_k, H_{m+1})$
12: ***end if***

3.4 Représentations condensées

Dans la présente section, nous présentons quelques représentations condensées des motifs fréquents dans des bases de données volumineuses. Une représentation condensée est une collection des motifs fréquents non redondants, à partir de laquelle, on peut dériver l'ensemble de tous les fréquents et leurs Supports. Dans les cas pratiques, le nombre d'éléments d'une représentation condensée est significativement plus petit que celui de l'ensemble de tous les motifs fréquents.

3.4.1 Sur les motifs fermés fréquents

Le problème de la fouille des règles d'association peut être entièrement traité dans le cadre de l'Analyse Formelle de Concepts (A.F.C.). L'A.F.C. fournit un cadre théorique fondamental pour plusieurs algorithmes de fouille des règles d'association. En A.F.C., un contexte formel est un triplet $\mathbb{K} = (\mathcal{E}, \mathcal{A}, \mathcal{R})$, où $\mathcal{E}$ et $\mathcal{A}$ sont des ensembles finis, et $\mathcal{R}$ est une relation binaire de $\mathcal{E}$ vers $\mathcal{A}$ [Wil82, GW99].

Définition 26: Considérons un contexte $\mathbb{K} = (\mathcal{E}, \mathcal{A}, \mathcal{R})$, où $\mathcal{E}$ et $\mathcal{A}$ sont des ensembles finis. Un tel contexte sera appelé contexte de la fouille de données. Les éléments de $\mathcal{E}$ seront appelés les *entités* et ceux de $\mathcal{A}$ les *attributs* ou *variables* du contexte $\mathbb{K}$.

La relation binaire $\mathcal{R}$ induit une correspondance de Galois entre les ensembles ordonnés $(\mathcal{P}(\mathcal{E}), \subseteq)$ et $(\mathcal{P}(\mathcal{A}), \subseteq)$ par le biais des fonctions f et g définies de la façon suivante [BM70] :

$$f : \mathcal{P}(\mathcal{E}) \to \mathcal{P}(\mathcal{A})$$
$$X \longmapsto f(X) = \bigcap_{x \in X} \{y \in \mathcal{A} : x\mathcal{R}y\} = \{y \in \mathcal{A} : \text{pour tout } x \in X, x\mathcal{R}y\}$$

$$g : \mathcal{P}(\mathcal{A}) \to \mathcal{P}(\mathcal{E})$$
$$Y \longmapsto g(Y) = \bigcap_{y \in Y} \{x \in \mathcal{E} : x\mathcal{R}y\} = \{x \in \mathcal{E} : \text{pour tout } y \in Y, x\mathcal{R}y\}$$

Ainsi, les applications $\varphi = f \circ g$ et $\varphi' = g \circ f$ sont respectivement des opérateurs de fermeture dans les deux ensembles ordonnés $(\mathcal{P}(\mathcal{A}), \subseteq)$ et $(\mathcal{P}(\mathcal{E}), \subseteq)$.

Soit $\mathbb{K}$ un contexte de la fouille de données. Dans la suite, nous utiliserons la correspondance de Galois (f, g) ainsi définie sur $\mathcal{P}(\mathcal{E})$ et $\mathcal{P}(\mathcal{A})$.

La proposition suivante caractérise le Support et la Confiance d'une règle d'association.

Proposition 16: Si X et Y deux motifs d'un contexte de la fouille de données $\mathbb{K}$, alors :

(i) $\text{Supp}(X) = \frac{|g(X)|}{|\mathcal{E}|}$;

(ii) $\text{Supp}(X \rightarrow Y) = \frac{|g(X \cup Y)|}{|\mathcal{E}|}$;

(iii) $\text{Conf}(X \rightarrow Y) = \frac{|g(X \cup Y)|}{|g(X)|}$.

Soit minsupp $\in [0, 1]$ un seuil minimum de Support. Notons par $\mathcal{F}$ l'ensemble de tous les motifs fréquents d'un contexte de la fouille de données $\mathbb{K}$. On a :

$$\mathcal{F} = \{X \subseteq \mathcal{A} : \text{Supp}(X) \geq \text{minsupp}\}.$$

Définition 27: Un motif X est dit *maximal fréquent* s'il est fréquent et que tous ses sur-ensembles sont infréquents. Formellement, l'ensemble M_F des motifs maximaux fréquents d'un contexte $\mathbb{K}$ est défini par :

$$M_F = \{X \subseteq \mathcal{A} : X \in \mathcal{F} \text{ et } \forall Y \supset X, Y \notin \mathcal{F}\}$$

Définition 28:

- Un motif X est dit *φ-fermé*, ou tout simplement *fermé*, si $\varphi(X) = X$. Il est dit *φ-fermé fréquent* s' il est à la fois φ-fermé et fréquent. Formellement, l'ensemble des motifs fermés fréquents d'un contexte $\mathbb{K}$ est défini par

$$\mathcal{F}_F = \{X \subseteq \mathcal{A} : \varphi(X) = X \text{ et } \text{Supp}(X) \geq \text{minsupp}\}.$$

- L'ensemble $M\mathcal{F}_F$ des motifs maximaux fermés fréquents est défini par

$$M\mathcal{F}_F = \{X \subseteq \mathcal{A} : X \in \mathcal{F}_F, \forall Y \supset X, Y \notin \mathcal{F}_F\}.$$

Remarque 12: Pour un motif $X \subseteq \mathcal{A}$, l'image $\varphi(X)$ sera appelée la *fermeture* de X. Elle correspond au plus petit fermé qui contient X.

Proposition 17: [PBTL99a] Pour un motif X, le Support de X est égal au Support de sa fermeture, *i.e.*,

$$\text{Supp}(\varphi(X)) = \text{Supp}(X).$$

Donc, la fermeture d'un motif fréquent est également fréquent.

Proposition 18: [PBTL99a] Les ensembles M_F des motifs maximaux fréquents et $M\mathcal{F}_F$ des motifs fermés maximaux fréquents sont identiques, *i.e.*, $M_F = M\mathcal{F}_F$.

Théorème 11: Soient $\mathbb{K}$ un contexte de la fouille de données et minsupp un seuil minimum de Support. L'ensemble $\mathcal{F}_F$ des motifs fermés fréquents est une représentation condensée des motifs fréquents, *i.e.*, c'est un sous-ensemble de motifs fréquents à partir duquel on peut dériver tous les motifs fréquents et leurs Supports.

Se basant sur les propriétés des motifs fermés fréquents, Pasquier et al. [PBTL99a] proposent un algorithme dénommé CLOSE pour générer tous les motifs fermés fréquents d'un contexte de fouille de données.

Comme dans l'algorithme APRIORI, les motifs sont ordonnés par un ordre lexicographique. Le pseudo-code pour générer les motifs fermés fréquents est donné dans l'algorithme 5. Les notations utilisées sont données dans le Tableau 3.2.

$\mathcal{FFC}_k$	Ensemble des k-groupes candidats des k-générateurs. Chaque élément possède 3 champs : générateur, fermé, Support
$\mathcal{FF}_k$	Ensemble de k-groupes fréquents des k-générateurs. Chaque élément possède 3 champs : générateur, fermé, Support

Tab. 3.2: *Notations utilisées dans l'algorithme 5*

CLOSE, présenté dans l'algorithme 5, est un algorithme itératif d'extraction des motifs fermés fréquents. À chaque itération, l'algorithme construit l'ensemble des motifs candidats ($\mathcal{FFC}_k$) (lignes 3 à 5). Cet ensemble est élagué par rapport à minsupp, obtenant ainsi l'ensemble des motifs fermés fréquents (lignes 6 à 9). Finalement, en utilisant cet ensemble, il construit l'ensemble des générateurs qui seront utilisés dans la prochaine itération (ligne 11). L'algorithme s'arrête quand la liste des générateurs est vide. Ainsi, chaque itération est composée de deux étapes.

(i) **Étape d'élagage :** durant cette étape, la procédure Gen-Closure est appliquée à chaque générateur $\mathcal{FFC}_k$ (ligne 5), déterminant ainsi son Support et sa fermeture. Notons ici une particularité de l'algorithme CLOSE qui élague par rapport à minsupp après avoir calculé les fermetures des générateurs dont certains peuvent être non fréquents.

(ii) **Étape de construction :** dans cette étape, on commence pour éliminer les générateurs non fréquents. Ensuite, la procédure Gen-Generator prend comme argument l'ensemble $\mathcal{FF}_k$ et calcule l'ensemble $\mathcal{FFC}_{k+1}$ contenant les $(k+1)$-motifs (ligne 11), qui seront utilisés dans l'itération suivante. À ce niveau, comme l'algorithme CLOSE dispose de l'ensemble des motifs fermés fréquents obtenus au niveau k, alors l'ensemble $\mathcal{FFC}_{k+1}$ est élagué comme suit. Pour tout p.générateur $\in \mathcal{FFC}_{k+1}$, si p.générateur est inclus dans la fermeture d'un des sous-ensembles, *i.e.* les éléments de $\mathcal{FF}_k$ dont la jointure a permis d'obtenir p.générateur. Dans ce cas p.générateur est éliminé dans $\mathcal{FFC}_{k+1}$.

Algorithme 5 (CLOSE):
Entrée : $\mathbb{K}$ *: contexte de fouille de données ;* minsupp *: seuil minimum support*
Sortie : $\mathcal{FF}$ *: ensemble des motifs fermés fréquents*

1: $\mathcal{FFC}_1.$*générateurs*$\leftarrow \{1\text{-}motifs\}$
2: ***for*** $(k = 1; \mathcal{FFC}_k.$*générateurs* $\neq \emptyset; k++$ $)$ ***do***
3: $\quad \mathcal{FFC}_k.$*fermés* $\leftarrow \emptyset$
4: $\quad \mathcal{FFC}_k.$*supports* $\leftarrow 0$
5: $\quad \mathcal{FFC}_k \leftarrow$ *Gen-Closure*$(\mathcal{FFC}_k)$
6: $\quad$ ***for all*** $c \in \mathcal{FFC}_k$ ***do***
7: $\quad\quad$ ***if*** $(c.$*support* $\geq$ minsupp$)$ ***then***
8: $\quad\quad\quad \mathcal{FF}_k \leftarrow \mathcal{FF}_k \cup \{c\}$
9: $\quad\quad$ ***end if***
10: $\quad$ ***end for***
11: $\quad \mathcal{FFC}_{k+1} \leftarrow$ *Gen-Generator*$(\mathcal{FF}_k)$
12: ***end for***
13: *Retourner* $\mathcal{FF} \leftarrow \bigcup_k \mathcal{FF}_k$

La procédure Gen-Closure qui prend comme arguments $\mathcal{FFC}_k$ et $\mathbb{K}$ contexte de la fouille de données génère les fermetures et calcule les supports des motifs candidats k-générateurs. Elle fonctionne de la façon suivante. Pour chaque

entité e, l'ensemble G_e est créé (ligne 2). G_e contient tous les générateurs $p \in \mathcal{FFC}_k$ qui sont sous-ensembles de $f\{e\}$. Chaque générateur $p \in G_e$, la fermeture associée ainsi que leur Support sont mis à jour (lignes 3 à 11). Si l'entité p.fermé est égal à l'ensemble vide, on affecte le p.fermé à $f\{e\}$ (lignes 4 à 5), sinon on affecte nouvelle fermeture p.fermé à l'intersection de l'ancien p.fermé et $f\{e\}$ (lignes 6 à 8). Ensuite, le Support p.support du motif p.fermé est incrémenté (ligne 9). Enfin, la procédure Gen-Closure retourne aux champs fermés et Supports de $\mathcal{FFC}_k$ mis à jour.

Algorithme 6 (Gen-Closure):
Entrée : $\mathcal{FFC}_k, \mathbb{K}$: *contexte de fouille de données*
Sortie : $\mathcal{FFC}_k.fermés, \mathcal{FFC}_k.supports$

1: ***for all*** $e \in \mathcal{E}$ ***do***
2: $G_e \leftarrow$ *Sous-Ensemble*$(\mathcal{FFC}_k.générateurs, f(\{e\}))$
3: ***for all*** $p.générateur \in G_e$ ***do***
4: ***if*** $p.fermé = \emptyset$ ***then***
5: $p.fermé \leftarrow f(\{e\})$
6: ***else***
7: $p.fermé \leftarrow p.fermé \cap f(\{e\})$
8: ***end if***
9: $p.support++$
10: ***end for***
11: ***end for***
12: *Retourner* $\bigcup\{p \in \mathcal{FFC}_k | p.fermé \neq \emptyset\}$

La procédure Gen-Generator qui prend comme argument $\mathcal{FF}_k$ génère l'ensemble $\mathcal{FFC}_{k+1}$ contenant les $(k+1)$-motifs. Cette procédure génère d'abord les $(k+1)$-générateurs candidats en joignant les k-générateurs de $\mathcal{FF}_k$ possédant les mêmes $(k-1)$-premiers attributs (ligne 1). Les $(k+1)$-générateurs candidats dont on sait qu'ils sont soit peu fréquents, soit non minimaux sont ensuite supprimés (lignes 2 à 8). Enfin, on supprime parmi les générateurs ceux dont la fermeture est déjà calculée (lignes 9 à 17).

Algorithme 7 (Gen-Generator):
Entrée : $\mathcal{FF}_k$
Sortie : $\mathcal{FFC}_{k+1}$

1: $\mathcal{FFC}_{k+1} \leftarrow$ *Apriori-Gen*$(\mathcal{FFC}_k.générateurs)$
2: ***for all*** $p.générateur \in \mathcal{FFC}_{k+1}.générateurs$ ***do***
3: ***for all*** $s \subseteq p.générateur$ $(s$: *k-motifs*$)$ ***do***

4: **if** $s \notin \mathcal{FF}_k.générateur$ **then**
5: $\mathcal{FFC}_{k+1} \leftarrow \mathcal{FFC}_{k+1} \setminus \{p\}$
6: **end if**
7: **end for**
8: **end for**
9: **for all** $p.générateur \in \mathcal{FFC}_{k+1}$ **do**
10: $S_p \leftarrow Sous\text{-}ensemble(\mathcal{FF}_k.générateurs, p.générateur)$
11: **for all** $s \in S_p$ **do**
12: **if** $(p.générateur \subseteq s.fermé)$ **then**
13: $\mathcal{FFC}_{k+1} \leftarrow \mathcal{FFC}_{k+1} \setminus \{p\}$
14: **end if**
15: **end for**
16: **end for**
17: *Retourner* $\mathcal{FFC}_{k+1}$

3.4.2 Sur les motifs libres

Le concept de motif *libre* a été introduit dans [BB00, Byk02, BBR03] et dans [BTP+00] sous le nom de motif *clé*.

Avant de présenter la notion de motifs libres, nous rappelons le concept de représentation *ϵ-adéquate* introduite dans [MT96]. Intuitivement, une représentation ϵ-adéquate est une représentation qui peut être substituée à une autre afin de répondre aux mêmes requêtes, plus efficacement, éventuellement au prix d'une erreur bornée par le paramètre ϵ.

Représentations ϵ-adéquates

Définition 29: Soit $\mathcal{S}$ une classe de structures. Soit $\mathcal{Q}$ une classe de requêtes définies sur $\mathcal{S}$. Considérons que la valeur d'une requête $Q \in \mathcal{Q}$ sur la structure $s \in \mathcal{S}$ soit un nombre réel de l'intervalle $[0, 1]$ noté $Q(s)$. Une *représentation ϵ-adéquate* pour $\mathcal{S}$ par rapport à $\mathcal{Q}$, est une classe de structures $\mathcal{C}$, une mise en correspondance $rep : \mathcal{S} \rightarrow \mathcal{C}$ et une fonction $m : \mathcal{Q} \times \mathcal{C} \rightarrow [0, 1]$ telle que $\forall Q \in \mathcal{Q}, \forall s \in \mathcal{S}, |Q(s) - m(Q, rep(s))| \leq \epsilon$. Autrement dit, lorsqu'on s'intéresse aux requêtes de fréquence, on cherche des représentations telles que l'erreur commise sur la fréquence calculée sur $rep(s)$ au lieu de s soit au plus ϵ pour tout s.

Exemple 8: Soit la classe de structures $\mathcal{DB}_{\mathcal{A}}$ de toutes les bases de données sur l'ensemble d'attributs $\mathcal{A}$. On considère $\mathcal{Q}_{\mathcal{A}}$ la collection de toutes les

requêtes qui retourne la fréquence d'un motif $X \subseteq \mathcal{A}$. Si l'on note Q_X une requête de $\mathcal{Q}_\mathcal{A}$ qui demande la fréquence d'un motif X alors $\mathcal{Q}_\mathcal{A} = \{Q_X | X \subseteq \mathcal{A}\}$ et la valeur de Q_X sur la base de données $\mathcal{E} \in \mathcal{DB}_\mathcal{A}$ est définie par $Q_X(\mathcal{E}) = \text{Supp}(X, \mathcal{E})$ (Support de X dans la base de données $\mathcal{E}$).

Un exemple de représentation ϵ-adéquate pour $\mathcal{DB}_\mathcal{A}$ par rapport à $\mathcal{Q}_\mathcal{A}$ est la représentation de $\mathcal{E} \in \mathcal{DB}_\mathcal{A}$ au moyen de $Freq\text{Supp}(\mathcal{E}, \epsilon)$. Ses composantes $rep, \mathcal{C}$ et m sont définies de la façon suivante : $\forall \mathcal{E} \in \mathcal{DB}_\mathcal{A}, rep(\mathcal{E}) = Freq\text{Supp}(\mathcal{E}, \epsilon)$, $\mathcal{C} = \{rep(\mathcal{E}) | \mathcal{E} \in \mathcal{DB}_\mathcal{A}\}$, $\forall Q_X \in \mathcal{Q}_\mathcal{A}$, $\forall c \in \mathcal{C}$, si $\exists (X, \alpha) \in rep(\mathcal{E})$ alors $m(Q_X, c) = \alpha$ sinon $m(Q_X, c) = 0$.

On vérifie que c'est une ϵ-adéquate pour $\mathcal{DB}_\mathcal{A}$ par rapport à $\mathcal{Q}_\mathcal{A}$ car $\forall a \in \mathcal{DB}_\mathcal{A}$, $|Q(a) - m(Q, rep(a))| \leq \epsilon$.

Nous nous intéressons à des représentations ϵ-adéquates qui ont une taille plus petite que la taille des structures initiales et nous parlons alors de représentations condensées.

Motifs δ-libres

Considérons un ensemble $\mathcal{A}$ d'attributs binaires. Un sous-ensemble X de $\mathcal{A}$ est dit un motif. Une transaction ou entité est formée par un sous-ensemble de $\mathcal{A}$; une base de données définie sur $\mathcal{A}$ est un ensemble de transactions de $\mathcal{A}$.

Définition 30: Soient $\delta \in [0, 1]$, X_1, X_2 deux motifs de $\mathcal{A}$ et $\mathcal{E}$ une base de données définie sur $\mathcal{A}$.

- La règle d'association $X_1 \rightarrow X_2$ est dite δ-*forte* basée sur $X = X_1 \cup X_2$ si $\text{Supp}(X_1) - \text{Supp}(X_1 \cup X_2) \leq \delta$.
- Un motif X est dit δ-*libre* si et seulement s'il n'existe aucune règle δ-forte basée sur X dans $\mathcal{E}$. L'ensemble de tous les motifs δ-libres de $\mathcal{E}$ sera noté $Free(\mathcal{E}, \delta)$ ou tout simplement $Free(\delta)$ si $\mathcal{E}$ est implicite dans le contexte.

Dans cette définition, δ est supposé avoir une valeur petite. Ainsi, une règle δ-forte est une règle avec très peu d'exceptions et donc une Confiance très élevée.

L'ensemble $Free(\mathcal{E}, \delta)$ vérifie la propriété d'anti-monotonicité.

Proposition 19: Soit X un motif de $\mathcal{A}$. Pour tout $Y \subseteq X$, si $X \in Free(\mathcal{E}, \delta)$ alors $Y \in Free(\mathcal{E}, \delta)$.

Le proposition suivante montre que le Support d'un motif peut être approché par un Support d'un motif libre.

Proposition 20: [BBR03] Soient $\mathcal{E}$ une base de données sur l'ensemble d'attributs $\mathcal{A}$, $X \subseteq \mathcal{A}$ et $\delta \in [0, 1]$, alors il existe $Y \subseteq X$ tel que $Y \in Free(\mathcal{E}, \delta)$ et $\text{Supp}(\mathcal{E}, Y) \geq \text{Supp}(\mathcal{E}, X) \geq \text{Supp}(\mathcal{E}, Y) - \delta|X|$

La Proposition 20 montre qu'on peut approcher le Support d'un motif X par le Support d'un motif Y δ-libre mais on ne peut pas déterminer l'ensemble Y. Le théorème suivant montre que l'ensemble Y peut être choisi parmi les motifs δ-libres sous-ensembles de X ayant Support minimal.

Théorème 12: [BBR03] Soient $\mathcal{E}$ une base de données définie sur l'ensemble d'attributs $\mathcal{A}$, $X \subseteq \mathcal{A}$ et $\delta \in [0, 1]$. On a : pour tout $Y \subseteq X$ tel que $Y \in Free(\mathcal{E}, \delta)$ et $\text{Supp}(\mathcal{E}, Y) = min(\{\text{Supp}(Z) | Z \subseteq X, Z \in Free(\mathcal{E}, \delta)\})$, alors $\text{Supp}(\mathcal{E}, Y) \geq \text{Supp}(\mathcal{E}, X) \geq \text{Supp}(\mathcal{E}, Y) - \delta|X|$.

Dans la pratique, le calcul de toutes les collections d'ensembles δ-libres est souvent délicat. Toutefois, on peut éviter ce calcul, du fait qu'une représentation ϵ-adéquate pour les réquêtes de fréquence, peut être obtenue à partir des ensembles δ-libres fréquents et la bordure négative correspondante.

Soit $\mathcal{E}$ une base de données définie sur l'ensemble d'attributs $\mathcal{A}$ et $\sigma \in [0, 1]$ un seuil minimum de Support.

Définition 31: L'ensemble des motifs δ-libres fréquents, noté $FreqFree(\mathcal{E}, \sigma, \delta)$, est défini par

$$FreqFree(\mathcal{E}, \sigma, \delta) = \{X \subseteq \mathcal{A} | X \in Free(\mathcal{E}, \delta) \; et \; \text{Supp}(X) \geq \sigma\}.$$

Définition 32: La bordure négative de $FreqFree(\mathcal{E}, \sigma, \delta)$, notée $\mathcal{B}d^-(\mathcal{E}, \sigma, \delta)$ est définie par

$$\mathcal{B}d^-(\mathcal{E}, \sigma, \delta) = \{X \subseteq \mathcal{A} | X \notin FreqFree(\mathcal{E}, \sigma, \delta) \; et \; \forall Y \subset X, Y \in FreqFree(\mathcal{E}, \sigma, \delta)\}$$

Ainsi, la bordure négative $\mathcal{B}d^-(\mathcal{E}, \sigma, \delta)$ est l'ensembles des motifs minimaux non σ-fréquents δ-libres. La technique d'approximation présentée dans [BBR00] n'utilise qu'un sous-ensemble de la bordure négative, noté $Free\mathcal{B}d^-(\mathcal{E}, \sigma, \delta)$, des motifs libres dans $\mathcal{B}d^-(\mathcal{E}, \sigma, \delta)$.

Définition 33: $Free\mathcal{B}d^{-}(\mathcal{E},\sigma,\delta) = \mathcal{B}d^{-}(\mathcal{E},\sigma,\delta) \bigcap Free(\mathcal{E},\delta)$

Nous avons besoin, pour construire la représentation ϵ-adéquate, des ensembles δ-libres et de leurs Supports. Les paires correspondantes sont définies de la façon suivante.

Définition 34: $FreqFree\text{Supp}(\mathcal{E},\sigma,\delta)$ est la collection des paires constituées d'un motif δ-libre σ-fréquent et leur Support ; formellement,

$$FreqFree\text{Supp}(\mathcal{E},\sigma,\delta) = \{(X, \text{Supp}(X)) | X \in FreqFree(\mathcal{E},\sigma,\delta)\}.$$

On peut définir une représentation ϵ-adéquate pour les requêtes de Support.

Définition 35: La représentation condensée des motifs σ-fréquents basée sur les motifs δ-libres (pour des valeurs de σ,δ et une classe de requêtes $\mathcal{Q}\subseteq\mathcal{Q}_{\mathcal{A}}$) est définie par une classe de structures $\mathcal{C}$, une mise en correspondance "rep" et une fonction "m", telle que $\forall\mathcal{E} \in \mathcal{DB}_{\mathcal{A}}$:

- $rep(\mathcal{E}) = (FreqFree\text{Supp}(\mathcal{E},\sigma,\delta), Free\mathcal{B}d^{-}(\mathcal{E},\sigma,\delta))$;
- $\mathcal{C} = \{rep(\mathcal{E}) | \mathcal{E} \in \mathcal{DB}_{\mathcal{A}}\}$;
- $\forall Q_X \in \mathcal{Q}, \forall c \in \mathcal{C}$, si $\exists Y \in Free\mathcal{B}d^{-}(\mathcal{E},\sigma,\delta)$, $Y\subseteq X$ alors $m(Q_X,c) = 0$ sinon $m(Q_X,c) = min(\{\alpha | \exists Z\subseteq X, (Z,\alpha) \in FreqFree\text{Supp}(\mathcal{E},\sigma,\delta)\})$.

Enfin, on peut établir que cette représentation est ϵ-adéquate pour les classes de bases de données et de requêtes définies ci-dessous.

Définition 36: $\mathcal{DB}_{\mathcal{A},s} = \{\mathcal{E} | \mathcal{E} \in \mathcal{DB}_{\mathcal{A}}\ et\ |\mathcal{E}| < s\}$ est la collection de bases de données qui n'ont pas plus de s lignes. $\mathcal{Q}_{\mathcal{A},n} = \{Q_X | X\subseteq\mathcal{A}\ et\ |X| \leq n\}$ est la collection des requêtes de Supports sur les motifs de taille au plus égale à n.

Théorème 13: [BBR03] La représentation des motifs σ-fréquents basée sur les motifs δ-libres (pour des valeurs σ,δ et la classe des requêtes $\mathcal{Q}_{\mathcal{A},n}$) est une représentation ϵ-adéquate pour $\mathcal{DB}_{\mathcal{A},s}$ par rapport à $\mathcal{Q}_{\mathcal{A},n}$ où $\epsilon = max(\sigma, \frac{n\delta}{s})$.

Algorithme de génération de motifs libres

Nous donnons maintenant un algorithme appelé MIN-EX présenté dans [BBR03] qui génère l'ensemble de tous les motifs δ-libres.

Algorithme 8 (MIN-EX):
Entrée : *$\mathcal{E}$ une base de données sur $\mathcal{A}$, σ, δ*
Sortie : *$FreqFree(\mathcal{E}, \sigma, \delta)$*
1: $\mathcal{C}_0 \leftarrow \{\emptyset\}$
2: $i \leftarrow 0$
3: ***while*** $\mathcal{C}_i \neq \emptyset$ ***do***
4: $\quad FreqFree_i \leftarrow \{X | X \in \mathcal{C}_i$ *et* X *σ-fréquent δ- libre*$\}$
5: $\quad \mathcal{C}_{i+1} \leftarrow \{X | X \subseteq \mathcal{A}$ *et* $\forall Y \subset X, Y \in \bigcup_{j \leq i} FreqFree_j\} \setminus \bigcup_{j \leq i} \mathcal{C}_j$
6: $\quad i \leftarrow i + 1$
7: ***end while***
8: *Retourner* $\bigcup_{j<i} FreqFree_j$

3.4.3 Sur les motifs essentiels

Nous présentons dans cette section une autre représentation condensée des motifs fréquents basée sur la notion de motifs *essentiels*. Elle est introduite dans [CCL05]. La notion de motifs essentiels est basée sur le principe d'identités d'inclusion-exclusion [Nar82]. Considérons un contexte de la fouille de données $\mathbb{K} = (\mathcal{E}, \mathcal{A}, \mathcal{R})$. Pour un motif X non vide de $\mathcal{A}$, notons $\overline{X}$ la négation de X.

Définition 37:
- Le *Support disjonctif de X, noté* $\text{Supp}(\vee X)$*, est défini par*

$$\text{Supp}(\vee X) = \frac{|\{E \in \mathcal{E} | X \cap E \neq \emptyset\}|}{|\mathcal{E}|}.$$

- *Le* Support *de $\overline{X}$ est défini par*

$$\text{Supp}(\overline{X}) = \frac{|\{E \in \mathcal{E} | X \cap E = \emptyset\}|}{|\mathcal{E}|}.$$

Remarque 13:
- Pour un motif X, les quantités $\text{Supp}(\vee X)$ et $\text{Supp}(\overline{X})$ sont des quantités complémentaires par rapport à 1, *i.e.*, $\text{Supp}(\vee X) + \text{Supp}(\overline{X}) = 1$.
- Le Support disjonctif d'un motif X est la proportion d'entités qui contiennent au moins un attribut de X.

	a_1	a_2	a_3	a_4	a_5
1	×	×	×	×	
2	×	×		×	
3			×		×
4	×		×	×	

Tab. 3.3: *Exemple d'un contexte de la fouille de données*

Exemple 9: Considérons le contexte de fouille de données présenté dans le Tableau 3.3. Nous avons : $\text{Supp}(a_1 \vee a_3) = 1$ et $\text{Supp}(\overline{a_1 a_3}) = 0$, $\text{Supp}(a_1 \vee a_2 \vee a_4) = \frac{3}{4}$ et $\text{Supp}(\overline{a_1 a_2 a_4}) = \frac{1}{4}$.

Les identités d'inclusion-exclusion, données dans la proposition suivante, permettent d'établir les relations entre le Support et le Support disjonctif pour un motif X.

Proposition 21: Soit X un motif. Nous avons les égalités suivantes :

$$\text{Supp}(X) = \sum_{Y \subseteq X, Y \neq \emptyset} (-1)^{|(Y)|-1} \text{Supp}(\vee Y) \tag{3.1}$$

$$\text{Supp}(\vee X) = \sum_{Y \subseteq X, Y \neq \emptyset} (-1)^{|(Y)|-1} \text{Supp}(Y) \tag{3.2}$$

Exemple 10: Dans le contexte présenté dans le Tableau 3.3, nous avons :

(1.) $\text{Supp}(a_1 a_3) = \text{Supp}(a_1) + \text{Supp}(a_3) - \text{Supp}(a_1 \vee a_3) = \frac{3}{4} + \frac{3}{4} - 1 = \frac{1}{2}$.

(2.) $\text{Supp}(a_1 \vee a_3) = \text{Supp}(a_1) + \text{Supp}(a_3) - \text{Supp}(a_1 a_3) = \frac{3}{4} + \frac{3}{4} - \frac{1}{2} = 1$

Définition 38: (**Motifs essentiels**)
Soient $\mathbb{K}$ un contexte de la fouille de données et X un motif de $\mathbb{K}$ tel que $X \neq \emptyset$.
On dit que X est un motif essentiel si et seulement si son Support disjonctif est différent des Supports disjonctifs de tous ses sous-ensembles directs, *i.e.*,

$$\text{Supp}(\vee X) \neq \max_{x \in X} \ (\text{Supp}(\vee(X \setminus \{x\}))).$$

Exemple 11: Pour le contexte présenté dans le Tableau 3.3 :

- le motif a_1a_3 est un motif essentiel. En effet, $\text{Supp}(a_1 \vee a_3) = 1 \neq \frac{3}{4} = \text{Supp}(a_1)$ et $\text{Supp}(a_1 \vee a_3) = 1 \neq \frac{3}{4} = \text{Supp}(a_3)$.
- le motif $a_1a_2a_4$ n'est pas un motif essentiel, car $\text{Supp}(a_1 \vee a_2 \vee a_4) = \frac{3}{4} = \text{Supp}(a_1 \vee a_2)$.

Notons M_E l'ensemble des motifs essentiels d'un contexte de la fouille de données et M_{EF} celui des motifs essentiels fréquents pour un seuil minsupp donné. La proposition suivante montre que la contrainte "X est essentiel" est anti-monotone.

Proposition 22: [CCL05] Soit X un motif d'un contexte $\mathbb{K}$ alors :

$$X \in M_E \text{ implique } [\text{ pour tout } Y \subseteq X, Y \in M_E].$$

Les deux propositions ci-dessous et le Théorème 14 montrent d'une part comment calculer le Support disjonctif à partir de motifs essentiels, d'autre part comment optimiser les identités d'inclusion-exclusion pour trouver le Support d'un motif fréquent.

Une méthode naïve pour calculer le Support d'un motif nécessite la connaissance de Support disjonctif de tous ses sous-ensembles. Cependant, il y a la possibilité de dériver le Support d'un motif en utilisant les Supports de motifs essentiels comme le montre le résultat de la proposition suivante.

Proposition 23: Soit X un motif d'un contexte $\mathbb{K}$, X qui n'est pas un motif essentiel. Nous avons :

$$\text{Supp}(\vee X) = \max_{Y \in M_E} (\{\text{Supp}(\vee Y) : Y \subseteq X\}). \tag{3.3}$$

Le résultat de la proposition suivante est une optimisation du calcul de Support d'un motif X, basée sur le concept de motifs essentiels.

Proposition 24: Soient X un motif et
$Z \in Argmax(\{\text{Supp}(\vee Y) : Y \subseteq X \text{ et } Y \in M_E\})$, on a :

$$\text{Supp}(X) = \sum_{Y \subseteq X, Y \neq \emptyset} (-1)^{|Y|-1} \begin{cases} \text{Supp}(\vee Z) & si \quad Z \subseteq Y \\ \text{Supp}(\vee Y) & \quad sinon \end{cases} \tag{3.4}$$

Le résultat du Théorème 14 est une méthode de dérivation de Support de X. Considérons minsupp $\in [0, 1]$ un seuil minimum de Support donné et notons $\mathcal{F}$ et $M_{EF} = M_E \cap \mathcal{F}$. Nous avons le théorème suivant.

Théorème 14: [CCL05] Soient
$X \in \mathcal{F}$ et $X \notin M_{EF}, Z \in Argmax(\{\text{Supp}(\vee Y) | Y \subseteq X \text{ et } Y \in M_{EF}\})$, alors on a :

$$\text{Supp}(X) = \sum_{Y \subseteq X, Y \neq \emptyset, Y \not\subseteq Z} (-1)^{|Y|-1} \text{Supp}(\vee Y) \tag{3.5}$$

L'ensemble des motifs essentiels n'est pas suffisant pour définir une représentation condensée des motifs fréquents car il ne permet pas de connaître si un motif est fréquent ou non. C'est la raison pour laquelle, on ajoute à l'ensemble des motifs essentiels la bordure positive des motifs fréquents pour tester si un motif quelconque est fréquent ou non. Si un motif est fréquent, le Théorème 14 permet de calculer le Support de ses conjonctions.

Définition 39: [CCL05] Soient $\mathbb{K} = (\mathcal{E}, \mathcal{A}, \mathcal{R})$ un contexte de la fouille de données et $\mathcal{F}$ l'ensemble des motifs fréquents (pour un seuil minsupp donné). On dit qu'un ensemble $\mathcal{C}$ est une *couverture* de $\mathcal{F}$ si le Support de chaque élément de $\mathcal{F}$ peut être retrouvé à partir des motifs de $\mathcal{C}$. De plus, si $\mathcal{C} \subseteq \mathcal{F}$, $\mathcal{C}$ est appelé *couverture parfaite* ou *représentation condensée* de l'ensemble $\mathcal{F}$.

Rappelons que la bordure positive des motifs fréquents [MT97], notée $Bd^+(\mathcal{F})$, est l'ensemble de motifs maximaux fréquents. Formellement, $Bd^+(\mathcal{F})$ est définie par

$$Bd^+(\mathcal{F}) = \{X \in \mathcal{F}; \text{pour tout } Y \supset X, Y \notin \mathcal{F}\}.$$

La représentation condensée des motifs fréquents basée sur le concept de motifs essentiels est obtenue en faisant la réunion des motifs essentiels fréquents et la bordure positive des motifs fréquents.

Théorème 15: [CCL05] Soient $Bd^+(\mathcal{F})$ la bordure positive des motifs fréquents pour un seuil minimum Support donné (minsupp) et M_{EF} l'ensemble des motifs essentiels fréquents. L'ensemble $Bd^+(\mathcal{F}) \cup M_{EF}$ est une couverture parfaite pour l'ensemble de motifs fréquents.

Exemple 12: Pour le contexte présenté dans le Tableau 3.3, considérons minsupp $= \frac{1}{2}$. Les motifs essentiels fréquents ainsi que leurs Supports disjonctifs sont donnés dans le Tableau 3.4. La bordure positive $Bd^+(\mathcal{F}) = \{a_1a_2a_4, a_1a_3a_4\}$.
Le motif $a_1a_2a_4$ est un motif fréquent car il appartient à $Bd^+(\mathcal{F})$. Calculons

Motif essentiel	Support disjonctif
a_1	$\frac{3}{4}$
a_2	$\frac{1}{2}$
a_3	$\frac{3}{4}$
a_4	$\frac{3}{4}$
a_1a_3	1
a_3a_4	1

Tab. 3.4: *Motifs essentiels fréquents et leurs Supports (*minsupp $= \frac{1}{2}$*)*

maintenant son Support et son Support disjonctif. Nous utilisons la Proposition 23 pour déterminer son Support disjonctif.
$\text{Supp}(a_1 \vee a_2 \vee a_4)) = max(\text{Supp}(\vee a_1), \text{Supp}(\vee a_2), \text{Supp}(\vee a_4)) = \text{Supp}(a_1) = \text{Supp}(a_4) = \frac{3}{4}$.
Appliquons le Théorème 14 pour calculer le Support de $a_1a_2a_4$.
Les motifs a_1 et a_4 sont des motifs essentiels fréquents inclus dans $a_1a_2a_4$. Ils correspondent au Support disjonctif maximal. Nous avons :
$Argmax(\{\text{Supp}(\vee Y) : Y \subseteq a_1a_2a_4 \text{ et } Y \in M_{EF}\}) = \{a_1, a_4\}$. Nous choisissons a_1 pour appliquer le Théorème 14. On a :$\text{Supp}(a_1a_2a_4) = \text{Supp}(a_1) + \text{Supp}(a_2) + \text{Supp}(a_4) - \text{Supp}(\vee(a_1a_2)) - \text{Supp}(a_1a_4) - \text{Supp}(a_2a_4) +$
$\text{Supp}(\vee(a_1a_2a_4)) = \text{Supp}(a_2) + \text{Supp}(a_4) - \text{Supp}(\vee(a_2a_4))$. Comme a_2a_4 n'est pas un motif essentiel, nous avons besoin de connaître le Support disjonctif de a_2a_4. Par application du Théorème 14, nous avons : $\text{Supp}(a_2) + \text{Supp}(a_4) - \text{Supp}(\vee(a_2a_4)) = \text{Supp}(a_2)$. Finalement, nous obtenons $\text{Supp}(a_1a_2a_4) = \frac{2}{4}$. Nous avons éliminé tous les motifs X tels que $a_1 \subseteq X \subseteq a_1a_2a_4$ dans la relation d'inclusion-exclusion car la somme de ses Supports disjonctifs, munis de coefficients (+1) et (-1), est nulle.

Pour générer la représentation condensée basée sur l'ensemble des motifs essentiels fréquents, Casali et al. [CCL05] proposent un algorithme par niveau, dénommé MEP (Mining Essential Patterns), qui utilise l'ensemble $Bd^+(\mathcal{F})$. Cet algorithme MEP utilise un algorithme Max-Set (par exemple : Max-Miner [Bay98], Gen-Max [GZ01], Pincer-search[LK98]) pour générer la bordure positive $Bd^+(\mathcal{F})$.

Algorithme 9 (MEP):
Entrée : $\mathbb{K} = (\mathcal{E}, \mathcal{A}, \mathcal{R})$, minsupp
Sortie : *Représentation condensée* $M_{EF} \cup Bd^+(\mathcal{E})$

1: $Bd^+(\mathcal{F}) \leftarrow$ *Max-Set*$(\mathcal{E})$
2: $L_1 \leftarrow \{$*1-motifs fréquents*$\}$
3: $i \leftarrow 1$
4: ***while*** $L_i \neq \emptyset$ ***do***
5: $\quad C_{i+1} \leftarrow Apriori - Gen(L_i)$
6: $\quad C_{i+1} \leftarrow \{X | X \in C_{i+1} | \exists Y \in Bd^+(\mathcal{F}) : X \subseteq Y\}$
7: $\quad$ *Calcul de* $\mathrm{Supp}(\vee X)$ *pour tout* $X \in C_{i+1}$
8: $\quad L_{i+1} \leftarrow \{X \in C_{i+1} | \nexists x \in X : \mathrm{Supp}(\vee X) = \mathrm{Supp}(\vee X \setminus \{x\})\}$
9: $\quad i \leftarrow i + 1$
10: ***end while***
11: *Retourner* $\bigcup_j L_j$

3.4.4 Sur les motifs non dérivables

Cette section concerne la représentation condensée des motifs fréquents basée sur la notion de motifs non dérivables (cf. Définition 41). Considérons un contexte de la fouille de données $\mathbb{K} = (\mathcal{E}, \mathcal{A}, \mathcal{R})$ où $\mathcal{E}$ et $\mathcal{A}$ sont des ensembles finis respectivement d'entités et d'attributs et $\mathcal{R}$ une relation binaire de $\mathcal{E}$ vers $\mathcal{A}$. La réunion de deux motifs X et Y sera notée XY, et l'ensemble $X \cup \{x, y\}$ sera noté Xxy.

Soit x un attribut de $\mathcal{A}$. On note $\overline{x}$ la négation de x. On dit qu'une entité $E \in \mathcal{E}$ contient $\overline{x}$ si elle ne contient pas x.

Définition 40: (i) Un *motif généralisé* est un ensemble formé d'attributs et négation d'attributs. On note $X\overline{Y}$ le motif généralisé $X \cup \{\overline{y} | y \in Y\}$.

(ii) On dit qu'une entité E contient un motif généralisé $G = X\overline{Y}$, et on note $G \subseteq E$, si $X \subseteq E$ et $E \cap Y = \emptyset$.

(iii) Le *Support* d'un motif généralisé, noté $\mathrm{Supp}(G)$, est défini par $\mathrm{Supp}(G) = \frac{|\{E \in \mathcal{E} | G \subseteq E\}|}{|\mathcal{E}|}$.

(iv) Un motif généralisé basé sur un motif I est un motif $G = X\overline{Y}$ tel que $I = X \cup Y$.

Exemple 13: Considérons la base de données présentée dans le Tableau 3.5. Il y a 8 motifs généralisés basés sur $a_1a_2a_3$: $a_1a_2a_3, a_1a_2\overline{a_3}, a_1\overline{a_2}a_3, \overline{a_1}a_2a_3, a_1\overline{a_2}\overline{a_3}, \overline{a_1}a_2\overline{a_3}, \overline{a_1}\overline{a_2}a_3$ et $\overline{a_1}\overline{a_2}\overline{a_3}$. Le motif généralisé $\overline{a_1}\overline{a_2}a_3$ a pour Support $\frac{1}{3}$ car e_2, e_3 contiennent a_3 et $e_2 \cap \{a_1a_2\} = \emptyset$ de même $e_3 \cap \{a_1a_2\} = \emptyset$.

$\mathcal{E} \setminus \mathcal{A}$	a_1	a_2	a_3	a_4	a_5
e_1	×	×	×	×	
e_2		×	×	×	
e_3	×		×	×	×
e_4	×	×	×		×
e_5		×		×	×
e_6	×			×	×

Tab. 3.5: *Contexte de la fouille de données*

Du principe d'inclusion-exclusion [GS97], nous avons le lemme suivant.

Lemme 1: [CG06] Soit $X\overline{Y}$ un motif généralisé basé sur I. On a :

$$\text{Supp}(X\overline{Y}) = \sum_{X \subseteq J \subseteq I} (-1)^{|J \setminus X|} \text{Supp}(J). \tag{3.6}$$

Comme le Support d'un motif généralisé est toujours positif ou nul, on a

$$\sum_{X \subseteq J \subseteq I} (-1)^{|J \setminus X|} \text{Supp}(J) \geq 0. \tag{3.7}$$

En isolant le Support de I, on obtient l'inégalité :

$$(-1)^{|I \setminus X|} \text{Supp}(I) \geq - \sum_{X \subseteq J \subset I} (-1)^{|J \setminus X|} \text{Supp}(J). \tag{3.8}$$

Comme $X \subseteq J \subset I$, on a : $|I \setminus J| = |I \setminus X| - |J \setminus X|$.
D'où l'équation,

$$-(-1)^{|I \setminus X|}(-1)^{|J \setminus X|} = (-1)^{|I \setminus X| + |J \setminus X| + 1} = (-1)^{|I \setminus J| + 1}. \tag{3.9}$$

(car $(-1)^{-|J \setminus X|} = (-1)^{-|J \setminus X|}(-1)^{2|J \setminus X|} = (-1)^{|J \setminus X|}$).
D'où le théorème :

Théorème 16: [CG06] Si $X \subseteq I \subseteq \mathcal{A}$ alors, on a :

$$\begin{aligned} \text{Supp}(I) &\leq \textstyle\sum_{X \subseteq J \subset I} (-1)^{|I \setminus J| + 1} \text{Supp}(J) \quad \text{si } |I \setminus X| \quad \text{impair ;} \\ \text{Supp}(I) &\geq \textstyle\sum_{X \subseteq J \subset I} (-1)^{|I \setminus J| + 1} \text{Supp}(J) \quad \text{si } |I \setminus X| \quad \text{pair.} \end{aligned} \tag{3.10}$$

On notera $\mathcal{R}_X(I)$ la formule (3.10) et on posera

$$\delta_X(I) = \sum_{X \subseteq J \subset I} (-1)^{|J \setminus X|} \text{Supp}(J). \tag{3.11}$$

Les différentes règles de dérivation de Support d'un motif peuvent être utilisées pour déterminer les bornes inférieure et supérieure d'un intervalle qui encadre le Support d'un motif I qu'on les notera respectivement par $LB(I)$ et $UB(I)$.

Le résultat de la proposition suivante se démontre en utilisant les formules (3.6) à (3.10).

Proposition 25: Soient X et I des motifs tels que $X \subseteq I$. On pose $Y = I \setminus X$.

$$|\text{Supp}(I) - \delta_X(I)| = \text{Supp}(X\overline{Y})$$

Le corollaire ci-après montre que ces limites sont non redondantes.

Corollaire 3: Aucune des règles $\mathcal{R}_X(I)$ n'est redondante. Pour tout motif $X \subseteq I$, il existe un contexte $\mathbb{K}' = (\mathcal{E}', \mathcal{A}, \mathcal{R})$, où $\mathcal{E}' \subseteq \mathcal{E}$, tel que $\mathcal{R}_X(I)$ donne l'unique meilleure approximation de Support de I.

Le théorème suivant montre que les règles $\mathcal{R}_X(I)$ sont complètes et correctes.

Théorème 17: [Cal03] Pour tout motif $I \subseteq \mathcal{A}$, les règles $\{\mathcal{R}_X(I) | X \subseteq I\}$ sont complètes et correctes pour déduire les meilleurs approximations des limites supérieure et inférieure de Support de I dans $\mathcal{E}$.

Du Théorème 17, il n'est pas possible de réduire la longueur de l'intervalle déduit de règles de dérivation. Toutefois, ceci ne signifie pas que toute valeur dans cet intervalle est un Support d'un motif. Le théorème suivant le confirme.

Théorème 18: [Cal03] Soient $\mathbb{K} = (\mathcal{E}, \mathcal{A}, \mathcal{R})$ un contexte de la fouille de données et I un motif. Pour tout nombre entier n sur l'intervalle $[LB(I), UB(I)]$, il existe un contexte $\mathbb{K}' = (\mathcal{E}', \mathcal{A}, \mathcal{R})$ tel que pour tout sous-ensemble J de I, $\text{Supp}(J, \mathbb{K}') = \text{Supp}(J, \mathbb{K})$ et $\text{Supp}(I, \mathbb{K}') = n$

Grâce aux règles de déduction présentées ci-dessus, il est possible de construire un sous-ensemble des motifs fréquents qui contient les mêmes informations que l'ensemble de tous ces motifs fréquents. Supposons que les règles

de déduction nous permettent de déduire exactement le Support d'un motif I, *i.e.*, $LB(I) = UB(I)$. Dans ce cas, il n'est plus nécessaire de faire le calcul explicite de Support de I (qui nécessite le parcours sur le contexte de la fouille de données). En effet, si nous voulons avoir le Support de I, il suffit de le dériver par les règles de déduction.

Définition 41: [CG06] Soit I un motif d'un contexte de la fouille de données. I est dit *motif dérivable* (DI) si le Support de I peut parfaitement être déterminé par les règles de déduction $\mathcal{R}_X(I)$. Tout motif qui n'est pas dérivable s'appelle *motif non dérivable* (NDI).

Se basant sur la notion de motifs non dérivables, T. Calders et B. Goethals [CG02, CG06] construisent une représentation condensée des motifs fréquents d'un contexte de la fouille de données. Nous établissons les propriétés de motifs non dérivables avant de présenter la représentation condensée basée sur cette notion.

Théorème 19: [CG06] Considérons un contexte de la fouille de données $\mathbb{K} = (\mathcal{E}, \mathcal{A}, \mathcal{R})$. Soient $I \subseteq \mathcal{A}$ un motif, "a" un attribut de $\mathcal{A}$. On a la relation suivante :

$$UB(Ia) - LB(Ia) \leq \frac{1}{2}(UB(I) - LB(I)).$$

Soit I un motif tel que $I = \{i_1, i_2, ..., i_n\}$. D'après le Théorème 19, on a :
$|UB(I) - LB(I)| \leq \frac{1}{2}(UB(I \setminus \{i_1\}) - LB(I \setminus \{i_1\})) \leq ...$
$\leq \frac{1}{2^{n-1}}(UB(\{i_n\}) - LB(\{i_n\})) = \frac{|\mathcal{E}|}{2^{n-1}}$.

Corollaire 4: Tout motif I tel que $|I| > log_2(|\mathcal{E}|) + 1$ est un motif dérivable.

Corollaire 5: (**Monotonicité**) Soient I, J des motifs tels que $J \subseteq I$. Si J est un motif dérivable alors I l'est aussi.

Corollaire 6: Si $\delta_X(I)$ est égal au Support de I alors tout sur-ensemble Ia de I sont des motifs dérivables, avec : $\text{Supp}(Ia) = \delta_X(Ia) = \delta_{Xa}(Ia)$.

La représentation condensée basée sur la notion de motifs non dérivables est donnée par le théorème suivant.

Théorème 20: [CG06] Soient $\mathbb{K} = (\mathcal{E}, \mathcal{A}, \mathcal{R})$ un contexte de la fouille de données et minsupp un seuil minimum de Support. L'ensemble $NDIRep$ défini par :

$$NDIRep(\text{minsupp}) = \{(I, \text{Supp}(I) | LB(I) \neq UB(I) \text{ et } \text{Supp}(I) \geq \text{minsupp}\}$$

est une représentation condensée des motifs fréquents.

Calders et Goethals [CG06] proposent un algorithme de génération des motifs fréquents non dérivables dénommé NDI-algorithm. Tableau 3.6 présente les notations utilisées dans l'algorithme NDI-Algorithm.

NDI-Rep	Ensemble de tous les motifs NDIs fréquents
C_l	Motifs candidats NDIs fréquents à l^{ieme} itération
F_l	Motifs NDIs fréquents à l^{ieme} itération
Gen	Ensemble des générateurs de pré-candidats NDIs
$PreC_{l+1}$	Motifs pré-candidats NDIs fréquents à l'itération l
$I.L; I.U$	Limites inférieure et supérieure de Support de motif I

Tab. 3.6: *Notations utilisées dans l'algorithme NDI-Algorithm*

Algorithme 10 (NDI-algorithm):
Entrée : *$\mathcal{E}$ une base de données sur l'ensemble d'attributs $\mathcal{A}$,* minsupp
Sortie : *Représentation condensée $NDIRep$*(minsupp)

1: $l \leftarrow 1; NDIRep \leftarrow \emptyset; C_1 \leftarrow \{1\text{-}motifs\}$
2: ***for all*** $I \in C_1$ ***do***
3: $I.L \leftarrow 0; I.U \leftarrow |\mathcal{E}|$
4: ***end for***
5: ***while*** $C_l \neq \emptyset$ ***do***
6: *Calcul de* Supp(X) *pour tout* $X \in C_l$
7: $F_l \leftarrow \{X \in C_l | \text{Supp}(X) \geq \text{minsupp}\}$
8: $NDIRep \leftarrow NDIRep \cup F_l$
9: $Gen \leftarrow \emptyset$
10: ***for all*** $I \in F_l$ ***do***
11: ***if*** Supp$(I) \neq I.L$ *et* Supp$(I) \neq I.U$ ***then***
12: $Gen \leftarrow Gen \cup \{I\}$
13: ***end if***
14: ***end for***
15: $PreC_{l+1} \leftarrow$ *Apriori-Gen(Gen)*
16: $C_{l+1} \leftarrow \emptyset$
17: ***for all*** $I \in PreC_{l+1}$ ***do***
18: *Calcul de intervalle limite* $[L, U]$ *de* I
19: ***if*** $l \neq u$ *et* $u \geq$ minsupp ***then***
20: $I.L \leftarrow l; I.U \leftarrow U; C_{l+1} \leftarrow C_{l+1} \cup \{I\}$
21: ***end if***

22: ***end for***
23: $l \leftarrow l + 1$
24: ***end while***
25: *Retourner* $NDIRep$

L'algorithme commence par initialiser l'ensemble des candidats à l'ensemble des singletons (ligne 1). Les étapes sur les lignes 2 à 4 affectent, pour chaque 1-motif I, l'intervalle $[LB(I), UB(I) = [0, |\mathcal{E}|]$. La boucle "while" (lignes 5 à 24) génère tous les ensembles de motifs fréquents NDIs. Elle s'arrête jusqu'à ce que l'ensemble des motifs candidats est vide. A la $l^{\text{ième}}$ itération de la boucle, les candidats de taille l sont calculés. Pour cela, on parcourt une fois la base de données (ligne 6). On affecte à F_l par la suite l'ensemble de ceux qui ont de Support supérieur à minsupp (ligne 7). L'ensemble F_l est l'ensemble des fréquents NDIs de taille l. Tous les élements de F_l sont insérés dans NDIRep (ligne 8). Dans les lignes 9 à 22, c'est la génération des nouveaux candidats. Pour tout motif de F_l tel que son Support est à la fois différent aux limites inférieure $I.L$ et supérieure $I.U$ est inséré dans l'ensemble Gen (ensemble des générateurs) (lignes 9 à 14). La procédure Apriori-Gen, appelée en ligne 15, qui prend comme argument l'ensemble Gen génère l'ensemble $PreC_{l+1}$ des précandidats de taille $l + 1$. Les étapes sur les lignes 16-22 concernent l'élagage de l'ensemble $PreC_{l+1}$.

3.5 Conclusion

Comme mentionné au début de ce chapitre, la fouille des règles d'association se fait en général en deux étapes :

- recherche des motifs fréquents ;
- dérivation des règles à partir de tous les motifs fréquents.

La résolution du second sous problème semble très aisée ; plus d'efforts ont été consacrés à la recherche des motifs fréquents. L'idée de trouver des représentations condensées de ces motifs a permis d'optimiser les recherches de motifs fréquents sans parcourir plusieurs fois la base transactionnelle. Nous avons présenté quelques méthodes pour trouver des représentations condensées des motifs fréquents, à savoir : les motifs fermés fréquents, les motifs libres, motifs essentiels et motifs non-dérivables. Une fois connue une représentation condensée des motifs fréquents, on dérive l'ensemble de

tous les motifs fréquents à partir d'une représentation condensée. Enfin, on génère l'ensemble de toutes les règles valides dans un contexte de la fouille de données. Remarquons toutefois que, le nombre des règles d'association valides au sens d'une mesure de qualité est souvent très élevé. Ce qui engendre un nouveau problème pour l'utilisateur, à savoir la difficulté de gérer les règles extraites. Pour pallier ce nouveau problème, d'autres alternatives ont été proposées, à savoir trouver un sous-ensemble des règles valides à partir duquel, on peut retrouver l'ensemble de toutes les règles d'association valides. Un tel sous-ensemble s'appelle *base* pour les règles d'association valides au sens d'une mesure de qualité. Une des contributions du présent travail concerne la génération des bases au sens de la mesure de qualité $\mathrm{M_{GK}}$.

4. MESURES DE QUALITÉ DES RÈGLES

4.1 Introduction

Les mesures de qualité ou mesures d'intérêt servent à évaluer, à classer les règles d'association d'un contexte de la fouille de données. Il existe plusieurs mesures de qualité proposées dans la littérature (on peut trouver des listes non exhaustives de ces mesures dans [GH06, HGB05b]). Les mesures les plus utilisées sont sans doute le Support et la Confiance [AIS93, AS94]. Toutefois, l'approche utilisant ces deux mesures présente un intérêt discutable pour l'utilisateur pour les raisons qui suivent. D'abord, les algorithmes utilisés pour générer les règles d'association d'un contexte binaire engendrent un très grand nombre de règles qui sont très difficiles à gérer et dont beaucoup n'ont que peu d'intérêt. Ensuite, la condition de Support qui est le moteur du processus d'extraction écarte les règles ayant un petit Support alors que certaines peuvent avoir une très forte Confiance et peuvent présenter un réel intérêt [Azé03a, Azé03b]. Enfin, l'utilisation exclusive des mesures de qualité Support et Confiance ne suffit pas pour garantir la qualité des règles détectées. En effet, comme le montre l'exemple du Tableau 4.1 (extrait de [LT04]), la règle $X \rightarrow Y$ possède un Support élevé (si l'on considère $n = 100$) et une Confiance élevée : $\text{Supp}(X \rightarrow Y) = 72$ et, $\text{Conf}(X \rightarrow Y) = 90$. Cependant, la Confiance de cette règle est égale à la probabilité $p(Y')$, soit $p(Y'|X') = p(Y')$; ce qui est la définition de l'indépendance statistique de X et Y et n'apporte aucune information nouvelle.

Pour pallier les faiblesses de ces deux mesures, plusieurs critères pour concevoir ou définir une bonne mesure de qualité ont été proposés et par la suite, différentes mesures ont été disponibles dans la littérature. Dans le présent chapitre, nous présentons différents critères pour définir une bonne mesure de qualité et nous étudions les comportements de certaines mesures

	X'	$\overline{X}'$	$\sum$
Y'	72	18	90
$\overline{Y}'$	8	2	10
$\sum$	80	20	100

Tab. 4.1: *Faiblesse de l'approche Support-Confiance*

face à ces critères. On constate qu'aucune de ces mesures ne vérifie simultanément ces critères. C'est la raison pour laquelle le problème de recherche de mesure de qualité à utiliser pour attraper les règles les plus pertinentes reste largement ouvert.

4.2 Définitions

Dans ce chapitre, nous nous plaçons dans le cadre d'un contexte de la fouille de données binaires. Rappelons qu'un contexte binaire est un triplet $\mathbb{K} = (\mathcal{E}, \mathcal{A}, \mathcal{R})$, où $\mathcal{E}$ est un ensemble fini d'entités et $\mathcal{A}$ un ensemble fini de variables ou attributs et $\mathcal{R}$ une relation binaire de $\mathcal{E}$ vers $\mathcal{A}$. Dans toute la suite, n désigne la cardinalité de $\mathcal{E}$ ($n = |\mathcal{E}|$). Rappelons que dans un contexte binaire, tout sous-ensemble X de $\mathcal{A}$ s'appelle motif. Avant de donner quelques définitions concernant les règles d'association et mesure de qualité des règles, nous posons les notations suivantes.

Notations

Soient X et Y deux motifs de $\mathcal{A}$. Notons :

- $X' = \{e \in \mathcal{E} | \forall x \in X, e\mathcal{R}x\}$, *i.e.*, l'ensemble de toutes les entités communes à tous les éléments de X : c'est le dual d'un motif X de $\mathcal{A}$, ou l'intension du motif X;
- $n_X = |X'|$: la cardinalité de X' ;
- $p(X) = \frac{n_X}{n}$: la proportion d'entités vérifiant X, estimation de la probabilité de X en considérant la probabilité discrète uniforme p sur l'espace probabilisable $(\mathcal{E}, \mathcal{P}(\mathcal{E}))$;
- $n_{X \cap Y} = |X' \cap Y'|$: la cardinalité de $X' \cap Y'$;
- $\overline{X}$: la négation de X., *i.e.* $\overline{X}(e) = 1$ si et seulement si il existe $x \in X$ tel que (non $e\mathcal{R}x$).

Dans la suite, pour un motif X de $\mathcal{A}$, nous appellerons $\overline{X}$ motif *négatif* tandis que X un motif *positif*. Le Tableau 4.2 présente un contexte binaire $\mathbb{K} = (\mathcal{E}, \mathcal{A}, \mathcal{R})$, où $\mathcal{E} = \{e_1, e_2, e_3, e_4, e_5\}$ et $\mathcal{A} = \{A, B, C, D, E\}$. Pour $X = \{B, C\}$, on a $X' = \{e_2, e_3, e_5\}$ et $\overline{X}' = \{e_1, e_4\}$. Plus souvent, les

$\mathcal{E} \setminus \mathcal{A}$	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. 4.2: *Contexte binaire*

règles d'association considérées sont des règles utilisant les motifs positifs. Certaines applications nécessitent non seulement la découverte des règles d'association de la forme $X \rightarrow Y$, mais aussi celle des règles de la forme $X \rightarrow \overline{Y}$. Ainsi, nous introduisons une définition plus générale des règles d'association utilisant les motifs positifs et ceux négatifs. Nous adoptons donc la définition suivante.

Définition 42:
- Une règle d'association d'un contexte binaire $\mathbb{K}$ est un couple (U, V), noté $U \rightarrow V$, où U et V sont des motifs positifs ou négatifs et $V \neq \emptyset$ (si V est un motif positif), $V \neq \overline{\mathcal{A}}$ (si V est un motif négatif).
- Pour une règle d'association $U \rightarrow V$, U et V sont appelés respectivement la *prémisse* et le *conséquent* de la règle.

Soient X et Y deux motifs positifs. Quatre types de règles d'association peuvent être obtenus à partir de X et Y :

(a) Une règle dite *positive* de la forme $X \rightarrow Y$ ou $Y \rightarrow X$;

(b) Une règle dite *négative à droite* de la forme $X \rightarrow \overline{Y}$ ou $Y \rightarrow \overline{X}$;

(c) Une règle dite *négative à gauche* de la forme $\overline{X} \rightarrow Y$ ou $\overline{Y} \rightarrow X$;

(d) Une règle dite *bilatéralement négative* de la forme $\overline{X} \rightarrow \overline{Y}$ ou $\overline{Y} \rightarrow \overline{X}$.

La validité d'une règle d'association est évaluée à partir d'une (ou de plusieurs) mesure(s) de qualité des règles. Nous donnons ci-dessous la définition d'une mesure de qualité des règles.

Définition 43: Une *mesure de qualité* ou *mesure d'intérêt* des règles est une fonction μ de l'ensemble des règles d'association à valeurs dans $\mathbb{R}$ telle que pour toute règle d'association $U \rightarrow V$, $\mu(U \rightarrow V)$ est fonction exclusivement de quatre paramètres $n, p(U'), p(V')$ et $p(U' \cap V')$, où p désigne la probabilité discrète uniforme sur l'espace probabilisable $(\mathcal{E}, \mathcal{P}(\mathcal{E}))$.

Par inspiration de logique formelle (où deux implications contraposées ont la même valeur logique), nous posons la définition suivante.

Définition 44: Une mesure de qualité μ sera dite *implicative* si pour toute règle d'association $X \rightarrow Y$, on a : $\mu(\overline{Y} \rightarrow \overline{X}) = \mu(X \rightarrow Y)$ [Tot03].

Définition 45: Une mesure de qualité des règles μ sera dite *symétrique* si pour toute règle d'association $X \rightarrow Y$, on a $\mu(Y \rightarrow X) = \mu(X \rightarrow Y)$. μ sera dite *parfaitement symétrique* si pour toute règle d'association $X \rightarrow Y$, on a $\mu(\overline{X} \rightarrow \overline{Y}) = \mu(X \rightarrow Y)$.

Avant de fournir quelques exemples de mesures de qualité des règles, nous présentons dans la section suivante des critères souhaités pour apprécier une mesure de qualité des règles.

4.3 Différents critères d'appréciation d'une mesure de qualité

Plusieurs mesures de qualité des règles ont été proposées dans la littérature. Pour analyser ces mesures, un certain nombre propriétés ont été souhaitées. Nous rappelons ci-dessous les principaux critères souhaités pour apprécier une mesure de qualité des règles.

4.3.1 Inteligibilité ou compréhensibilité

(P1) Une mesure doit être intelligible [LMV+04, LT04], *i.e.*, elle doit avoir un sens "concret" qui soit parlant à l'utilisateur, elle doit être facile à interpréter. C'est le cas des mesures Support et Confiance, elles ont un sens "concret". Elles sont facilement interprétables par l'utilisateur. Considérons deux règles d'association $X_1 \rightarrow Y_1$ et $X_1 \rightarrow Y_2$ ayant le même support et telles que $\text{Conf}(X_1 \rightarrow Y_1) = 2 \times \text{Conf}(X_2 \rightarrow Y_2)$. La seule connaissance de la Confiance permet à l'utilisateur de savoir que la règle $X_1 \rightarrow Y_1$ est deux fois plus fiable que la règle $X_2 \rightarrow Y_2$.

4.3.2 Nature des règles ciblées par la mesure

Une mesure doit distinguer les différentes règles obtenues par la combinaison de X et Y.

(P2) Une mesure doit impérativement permettre de choisir entre $X \rightarrow Y$ et $X \rightarrow \overline{Y}$ [Fre99].

(P3) On préfère les mesures non symétriques qui respectent la nature des règles d'association "si X, alors Y" [LT04, LMV$^+$04].

(P4) Une mesure doit évaluer de la même façon $X \rightarrow Y$ et $\overline{Y} \rightarrow \overline{X}$, *i.e.*, elle doit être implicative [Kod99, DRT07].

4.3.3 Sensibilité à l'apparition des contre-exemples

(P5) L'évaluation de l'intérêt d'une règle peut se mesurer favorablement en fonction du nombre élevé d'exemples de la règle ou en fonction du nombre faible de ses contre-exemples [Fre99].

4.3.4 Situations de référence

L'utilisation de mesures de qualité prenant des valeurs positives pour les règles intéressantes permet de se rapprocher des *a priori* de l'utilisateur sur la notion de qualité.
Selon Piatetstky-Shapiro [PS91], une bonne mesure μ de qualité de la règle $X \rightarrow Y$ doit être :

(P6) $\mu(X \rightarrow Y) < 0$ en cas de répulsion entre la prémisse et le conséquent d'une règle, *i.e.*, $p(Y'|X') < p(Y')$;

(P7) $\mu(X \rightarrow Y) = 0$ en cas de l'indépendance entre la prémisse et le conséquent d'une règle, *i.e.*, $p(Y'|X') = p(Y')$;

(P8) $\mu(X \rightarrow Y) > 0$ en cas d'attraction entre la prémisse et le conséquent d'une règle, *i.e.*, $p(Y'|X') > p(Y')$.

4.3.5 Variation non linéaire par rapport à $p(X \cap \overline{Y})$ au voisinage de 0^+

(P9) Pour certains auteurs [GKCG01], il est souhaitable qu'une mesure μ ait une décroissance faible au voisinage de règle logique (*i.e.*, $p(X' \cap \overline{Y}')$ tends vers 0) plutôt que décroissance rapide ou linéaire. Ceci reflète le fait que l'utilisateur peut tolérer peu de contre-exemples tout en conservant l'intérêt d'une règle.

4.3.6 Impact de la rareté du conséquent

(P10) Une mesure μ doit être une fonction croissante de $1 - p(Y')$, *i.e.*, la rareté du conséquent. En effet, plus le conséquent Y est rare, plus le fait qu'il contienne la prémisse X pour une modélisation donnée est intéressant.

4.3.7 Sensibilité à la taille de données

(P11) Une mesure est dite *descriptive*, si elle ne change pas en cas de dilatation des données, dans le cas contraire, elle est dite mesure *statistique* [LT04]. Donc, pour une mesure statistique μ, la taille de données n doit intervenir dans son évaluation [LT04]. Pour une mesure statistique, en fixant les quantités marginales $p(X')$ et $p(Y')$, il est intéressant de savoir comment évaluer la règle $X \rightarrow Y$ si on augmente la taille de données n. Si une mesure varie de façon croissante avec n et admet une valeur maximale, alors elle risque de perdre son pouvoir discriminant quand n devient suffisamment grand.

4.3.8 Fixation d'un seuil

(P12) Les mesures de qualité retenues pour extraire les règles d'association doivent pouvoir être utilisées avec un seuil d'élagage de manière à éliminer toutes les règles qui n'intéressent pas l'utilisateur [LMV+04]. Les mesures ayant un sens concret pour l'utilisateur, ainsi que les mesures normalisées et ayant un caractère statistique se prête bien à la détermination d'un seuil d'élagage. Ce seuil peut être fixé par l'utilisateur soit avant la phase d'extraction des règles d'association, soit lors d'une phase de post-élagage des règles. Néanmoins, lorsque le seuil est déterminé a priori par l'utilisateur, ce seuil ne prend pas en considération la nature des données et peut conduire à des résultats ne présentant pas toujours les données. L'utilisation de seuils

d'élagage calculés directement à partir des données peut permettre d'éviter ce problème : il est donc souhaitable que le seuil soit statistique. De tels seuils peuvent être obtenus à partir des valeurs moyennes observées sur les données. Une méthode classique en fouille de données [Ler84] consiste à centrer et réduire les valeurs observées.

4.3.9 Déviation à l'équilbre

(P13) Une mesure de qualité doit tenir compte de l'équilibre, *i.e.*, lorsque les nombres d'exemples et de contre-exemples de la règle sont égaux, une mesure de qualité doit avoir une valeur constante, ou tout au moins asymptotiquement constante en fonction de la taille de l'échantillon [BGBG05].

4.3.10 Comportement par rapport à la taille de la prémisse ou du conséquent

Freitas propose dans [Fre99] un critère pour apprécier une mesure de qualité : (P14) une bonne mesure de qualité doit être une fonction décroissante de la taille de la prémisse (resp. du conséquent) lorsque les autres paramètres sont fixés.

4.4 Quelques exemples de mesures de qualité

Nous donnons dans cette section quelques exemples de mesures de qualité des règles tout en essayant de donner de sens concret à ces différentes mesures. Pour cela, considérons une règle d'association $X \rightarrow Y$ d'un contexte de fouille de données $\mathbb{K}$.

4.4.1 Support

Le Support [AIS93] de $X \rightarrow Y$ est défini par

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

Il indique la proportion d'entités vérifiant à la fois la prémisse et le conséquent de la règle. Il est une mesure symétrique, non implicative. La mesure Support souvent utilisée en la fouille des règles d'association sert à élaguer les règles non intéressantes du fait de sa propriété d'antimonotonicité. Il prend ses valeurs sur l'intervalle $[0, 1]$.

4.4.2 Confiance

La Confiance [AIS93] de $X \rightarrow Y$ est définie par

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

Elle indique la proportion d'entités vérifiant le conséquent parmi celles vérifiant la prémisse de la règle. Cette mesure est non symétrique et non implicative. Elle prend la valeur de référence $\frac{1}{2}$ à l'équilibre. Elle n'est pas sensible à la taille de données : c'est donc une mesure descriptive. Elle prend ses valeurs sur l'intervalle $[0, 1]$.

4.4.3 Rappel

La mesure rappel [LFZ99] est définie par

$$\mathit{Rappel}(X \rightarrow Y) = p(X'|Y') = \frac{p(X' \cap Y')}{p(Y')}.$$

Elle permet d'évaluer la proportion d'entités vérifiant la prémisse parmi celles vérifiant le conséquent de la règle. Elle est évidemment non symétrique, non implicative et insensible à la taille de données. Rappel est une mesure descriptive. La mesure de qualité Rappel prend ses valeurs sur l'intervalle $[0, 1]$.

4.4.4 Lift

La mesure Lift [BMS97] est définie par

$$Lift(X \rightarrow Y) = \frac{p(X' \cap Y')}{p(X')p(Y')}.$$

Cette mesure de qualité représente le rapport d'indépendance entre la prémisse et le conséquent de la règle. Lift est une mesure symétrique non implicative. Il est sensible à la taille de données : c'est une mesure statistique. Lift prend ses valeurs sur $[0, +\infty[$

4.4.5 Conviction

La mesure conviction [BMS97] est défine par

$$Conviction(X \rightarrow Y) = \frac{p(X')p(\overline{Y}')}{p(X' \cap \overline{Y}')}.$$

Une valeur de Conviction élevée indique que le nombre de contre-exemples de la règle est inférieur à celui attendu sous l'hypothèse d'indépendance entre la prémisse et le conséquent de la règle. Elle est une mesure non symétrique, mais implicative, contrairement à la mesure $Lift$. La Conviction prend ses valeurs sur $[0, +\infty]$

4.4.6 Pearl

La mesure de Pearl [Pea88] est définie par

$$Pearl(X \rightarrow Y) = p(X')|p(Y'|X') - p(Y')|.$$

Cette mesure permet d'évaluer l'intérêt d'une règle par rapport à l'hypothèse de l'indépendance entre la prémisse et le conséquent d'une règle d'association. Elle est une mesure symétrique, implicative et sensible à la taille de données. La mesure de Pearl est une mesure statistique. Elle prend ses valeurs sur l'intervalle $[0, 1]$.

4.4.7 ϕ-coefficient

La mesure ϕ-coefficient [LGR81] est définie par

$$\phi(X \rightarrow Y) = \frac{p(X' \cap Y') - p(X')p(Y')}{\sqrt{p(X')p(Y')p(\overline{X}')p(\overline{Y}')}}.$$

Elle évalue l'écart à l'indépendance et est normalisée par le produit des marges du tableau de contingence obtenu en faisant le croisement de X' et Y'. Elle est une mesure symétrique, non implicative et sensible à la taille de données : c'est une mesure statistique. Elle prend ses valeurs sur l'intervalle $[-1, 1]$.

4.4.8 Pietetsky-Shapiro

La mesure de Piatetsky-Shapiro [PS91] est définie par :

$$Piatetsky(X \rightarrow Y) = np(X')(p(Y'|X') - p(Y')).$$

La mesure de Piatetsky-Shapiro évalue l'intérêt d'une règle par rapport à son écart à l'indépendance. Elle est une mesure symétrique, sensible à la taille de données mais elle n'est pas une mesure de qualité implicative. Elle prend ses valeurs sur l'intevalle $[-n, n]$.

4.4.9 Nouveauté

La Nouveauté [LFZ99] est définie par

$$\text{Nouveauté}(X \rightarrow Y) = p(X' \cap Y') - p(X')p(Y').$$

Comme la mesure de Piatetsky-Shapiro, la Nouveauté est une mesure de l'écart à l'indépendance entre la prémisse et le conséquent de la règle. Elle est symétrique donc non implicative. Elle dépend de la taille de données. Elle prend ses valeurs sur l'intervalle $[-1, 1]$.

4.4.10 Confiance centrée

La Confiance centrée [LT04] est définie par

$$\text{Conf}_{\text{centrée}} = p(Y'|X') - p(Y').$$

Elle permet de prendre en considération la taille du conséquent de la règle. Cette mesure n'est ni symétrique ni implicative, mais elle est sensible à la taille de données. La mesure Confiance centrée permet de mesurer l'influence de réalisation du conséquent par apport à celle de la prémisse d'une règle. La valeur de Confiance centrée voisin de zéro implique que la règle proche de la situation d'indépendance entre la prémisse et le conséquent. Cette mesure est aussi appelée parfois Valeur ajoutée (Added value) (cf. [GH06]). Elle prend ses valeurs sur l'intervalle $[-1, 1]$.

4.4.11 Loevinger

La mesure de Loevinger [Loe47] est définie par

$$Loevinger(X \rightarrow Y) = \frac{p(Y'|X') - p(Y')}{p(\overline{Y}')}.$$

La mesure de Loevinger est une mesure non symétrique, implicative et sensible à la taille de données. Cette mesure normalise la mesure Confiance centrée par le nombre d'entités ne vérifiant pas le conséquent de la règle. Elle prend ses valeurs sur l'intervalle $]-\infty, 1]$. Remarquons que cette mesure de Loevinger est identique à la mesure Satisfaction introduite dans [LFZ99]. Elle est définie par

$$\text{Satisfaction}(X \rightarrow Y) = \frac{p(\overline{Y}') - p(\overline{Y}'|X)}{p(\overline{Y}')}.$$

4.4.12 Moindre contradiction

La mesure moindre contradiction [Azé03b, AK02] est définie par

$$Contramin(X \rightarrow Y) = \frac{p(X' \cap Y') - p(X' \cap \overline{Y}')}{p(Y')}.$$

Cette mesure évalue la différence entre le nombre d'exemples de contre-exemples d'une règle. Cette différence est normalisée par le nombre d'entités vérifiant le conséquent de la règle. Elle permet de sélectionner des règles possédant plus d'exemples que de contre exemples. Elle prend ses valeurs sur l'intervalle $]-\infty, +\infty[$.

4.4.13 Sebag-Schœnauer

La mesure de Sebag-Schœnauer [SS88] est définie par

$$Sebag(X \rightarrow Y) = \frac{p(X' \cap Y')}{p(X' \cap \overline{Y}')}.$$

Cette mesure évalue le rapport entre le nombre d'exemples et de contre-exemples de la règle. À l'équilibre elle prend la valeur 1. Si la valeur de mesure est supérieure à 1, la règle possède plus d'exemples que de contre-exemples. Elle n'est ni symétrique, ni implicative. Elle prend ses valeurs sur $[0, +\infty]$.

4.4.14 Indice d'implication

L'indice d'implication [LGR81] est défini par

$$Ind\text{-}Implication(X \rightarrow Y) = \sqrt{n}\frac{p(X' \cap Y') - p(X')p(Y')}{\sqrt{p(X')p(Y')}}$$

L'indice d'implication permet d'évaluer la petitesse du nombre de contre-exemples à la règle $X \rightarrow Y$ par rapport à la quantité attendue sous l'hypothèse d'indépendance. Elle est une mesure non symétrique, non implicative, mais elle varie en fonction de la taille de données : c'est une mesure statistique. Elle prend ses valeurs sur l'intervalle $[-\sqrt{n}, +\infty[$.

4.4.15 J-mesure

La mesure J-mesure [GS88] est définie par

$$J\text{-mesure}(X \rightarrow Y) = p(X' \cap Y')log(\frac{p(X' \cap Y')}{p(X')p(Y')}) + p(X' \cap \overline{Y}')log(\frac{p(X' \cap \overline{Y}')}{p(X')p(\overline{Y}')})$$

Cette mesure trouvant son origine dans le domaine de la théorie d'information permet d'estimer la quantité d'information apportée par l'étude de la règle $X \rightarrow Y$. Cette mesure n'est ni symétrique, ni implicative.

Les Tableaux 4.3 et 4.4 présentent les comportements de ces différentes mesures face aux critères souhaités pour une bonne mesure de qualié des règles d'association. Lecture des tableaux : O signifie que le critère est vérifié par la mesure et N signifie que le critère n'est pas vérifié par la mesure. Par exemple, Le Support vérifie le critère (P1) et il ne vérifie pas (P2).

4.5 Etudes comparatives des mesures de qualité

Nous avons vu précédemment que les mesures de qualité des règles ont des comportements différents face aux critères d'appréciation d'une mesure de qualité des règles. Des études comparatives ont été menées pour apporter un nouvel éclairage sur les mesures de qualité existantes.

4.5.1 Transformation affine de la Confiance

Des études comparatives faites par Lallich et Teytaud [LT04] sur les mesures de qualité apportent un éclairage intéressant sur les différentes mesures pro-

Mesure	P1	P2	P3	P4	P5	P6	P7	P8
Support	O	N	N	N	O	N	N	O
Confiance	O	N	O	N	O	N	N	O
Rappel	O	N	O	N	O	N	N	O
Lift	O	O	N	N	O	N	N	O
Pearl	N	N	N	O	O	O	O	O
ϕ-coefficient	O	O	N	N	O	O	O	O
Piatetsky-Shapiro	N	O	N	N	O	O	O	O
Nouveauté	O	O	N	O	O	O	O	O
Confiance-Centrée	N	O	O	N	O	O	O	O
Loevinger	N	O	O	O	O	O	O	O
Moindre Contradiction	O	N	O	N	O	N	N	N
Sebag-Shoenauer	O	N	O	N	O	N	N	O
Conviction	O	O	O	O	O	N	N	O
Ind-Implication	N	O	N	N	O	O	O	O
J-mesure	N	N	O	N	N	O	O	O

Tab. 4.3: *Propriétés des mesures de qualité des règles*

posées. La plupart des mesures de qualité peuvent s'exprimer comme transformation affine de la Confiance. Rappelons qu'une mesure de qualité μ est une transformation affine de la Confiance si elle peut s'exprimer sous la forme suivante :

$$\mu(X \rightarrow Y) = a(p(Y'|X') - b)$$

avec "a" et "b" deux paramètres qui ne dépendent que des paramètres $p(X'), p(Y')$ et éventuellement de n. Ces mesures peuvent s'interpréter comme centrage et réduction de la Confiance. La plupart de ces mesures sont centrées sur $p(Y')$.

Le centrage de la Confiance par rapport à $p(Y')$ permet de corriger un des défauts de la Confiance, en comparant la valeur observée à celle attendue sous l'hypothèse d'indépendance. Le changement d'échelle diffère suivant les mesures et suivant les buts recherchés. De plus, ce changement d'échelle permet de différencier les mesures centrées sur $p(Y')$. Comme mentionné dans [LT04], ces mesures corrigent la principale critique faite à la Confiance, mais elles héritent, par construction, les différentes caractéristiques de la Confiance. Ainsi, si les quantités $p(X')$ et $p(Y')$ sont fixées, alors ces mesures sont

Mesure	P9	P10	P11	P12	P13	P14
Support	O	N	O	N	N	N
Confiance	N	N	N	N	N	N
Rappel	O	N	N	N	N	O
Lift	O	O	O	N	N	O
Pearl	O	O	N	N	N	N
ϕ-coefficient	O	N	O	N	N	N
Piatetsky-Shapiro	O	O	O	N	N	O
Nouveauté	N	O	O	O	N	O
Confiance-Centrée	O	O	O	N	N	O
Loevinger	O	O	O	N	N	O
Moindre Contradiction	O	N	N	N	O	O
Sebag-Shoenauer	O	N	N	N	O	O
Conviction	O	O	O	N	N	N
Ind-Implication	O	O	O	N	N	O
J-mesure	N	N	O	N	N	N

Tab. 4.4: *Propriétés des mesures de qualité des règles (suite)*

une fonction affine du nombre de contre-exemples, donc varie linérairement par rapport à cette valeur. D'autre part, toutes ces mesures à l'exception de la mesure Piatetsky-Shapiro et de l'indice d'implication ne dépendent pas de n et elles sont donc invariantes par rapport à la variation de la taille de données. Nous donnons ci-joint la liste de ces mesures : Confiance centrée, Lift, Piatetsky-Shairo, Loevinger, Indice d'Implication, Pearl, la moindre contradiction, la nouveauté, la satisfaction.

4.5.2 La nouveauté

Lavrac et al. [LFZ99] présentent une étude nouvelle sur quelques mesures de qualité. Ces mesures sont transformables dont la transformée commune est la mesure de qualité Nouveauté. Ils proposent de définir la Confiance relative, la Sensibilité relative, la Spécificité relative et la fiabilité negative relative. Les mesures relatives sont obtenues à partir des mesures initialement définies en enlevant à la valeur attendue sous l'hypothèse d'indépendance. Ainsi, nous avons le Tableau 4.5 présentant les mesures nouvellement définies. A partir

Mesure relative	Expression
Confiance relative	$p(Y'\|X') - p(Y')$
Fiabilité negative relative	$p(\overline{Y}'\|\overline{X}') - p(\overline{Y}')$
Spécificité relative	$p(X'\|Y') - p(X')$
Sensibilité relative	$p(\overline{X}'\|\overline{Y}') - p(\overline{X}')$

Tab. 4.5: *Tableau des mesures relatives*

des mesures relatives ainsi obtenues, on effectue une nouvelle transformation. Pour la mesure Confiance relative, par exemple, on la multiplie avec $p(X')$. Cette nouvelle transformation a été conçue pour pallier un des défauts de la Confiance relative, à savoir trouver les règles ayant un petit Support et une Confiance relative très élevée. Les autres mesures se transforment de la même manière. Ainsi, nous avons le Tableau 4.6 présentant les mesures pondérées relatives finalement obtenues.

Les auteurs montrent que ces différentes mesures pondérées sont identiques. Elles sont égales à la mesure Nouveauté.

Mesure relative	Expression
Confiance relative pondérée	$p(X')(p(Y'\|X') - p(Y'))$
Fiabilité negative relative pondérée	$p(\overline{X}')(p(\overline{Y}'\|\overline{X}') - p(\overline{Y}'))$
Spécificité relative pondérée	$p(Y')p(X'\|Y') - p(X'))$
Sensibilité relative pondérée	$p(\overline{Y}')p(\overline{X}'\|\overline{Y}') - p(\overline{X}'))$

Tab. 4.6: *Tableau des mesures relatives pondérées*

4.6 Conclusion

La recherche de qualité des connaissances dans les bases de données est un problème d'actualité pour les chercheurs travaillant dans le domaine de la fouille de données, en particulier dans la fouille des règles d'association. Les règles d'association sont des implications conditionnelles entre les attributs. Les mesures de qualité, comme leurs noms l'indiquent, servent à évaluer ou à valider les règles d'association. Nous avons présenté dans ce chapitre plusieurs critères souhaités pour une bonne mesure de qualité des règles d'association et présenté par la suite quelques mesures proposées dans la littérature. Nous constatons qu'aucune des mesures proposées ne vérifie simultanément ces différents critères. Nous préférons avoir une mesure de qualité qui vérifie plusieurs critères d'appréciation parmi ceux proposés. Nous utiliserons cette mesure pour attraper les règles intéressantes dans des bases de données. Par ailleurs, les études comparatives des mesures de qualité présentées dans [LT04, LFZ99] que nous avons présentées ci-dessus ne regroupent que des petits nombres des mesures. Nous proposerons plus tard une approche analytique dénommée *normalisation* permettant de regrouper la plupart des mesures proposées dans la littérature.

5. LA MESURE DE QUALITÉ DE GUILLAUME-KHENCHAFF : $\mathrm{M_{GK}}$

5.1 Introduction

Nous avons vu au chapitre précédent des différents critères pour apprécier une mesure de qualité des règles d'association. Il est facile de vérifier qu'aucune des mesures disponibles dans la littérature ne satisfait simultanément à ces différents critères souhaitésqui sont stipulés dans le chapitre 4. Se pose alors la question "quelle mesure choisir ?". Dans le présent chapitre, nous présentons des études des propriétés de la mesure de Guillaume-Khenchaff ($\mathrm{M_{GK}}$). Les travaux de S. Guillaume ont été inspirés par la mesure de Loevinger avec sa capacité d'identifier les zones d'attraction et de répulsion, et guidée par la volonté de pallier les inconvénients de la mesure Confiance, notamment l'inconvenient de sélectionner des règles situées dans la zone de répulsion entre la prémisse X et le conséquent Y d'une règle d'association $X \rightarrow Y$, pour définir $\mathrm{M_{GK}}$ [Gui00]. Elle a insisté sur les propriétés de $\mathrm{M_{GK}}$ du fait qu'elle vérifie les principes de Piatetsky-Shapiro [PS91] sans toutefois étudier de façon approfondie ses propriétés mathématiques. Indépendamment de S. Guillaume, Wu et al. ont introduit dans [WZZ04] cette même mesure dénommée $CPIR$ (Conditional Probability Increment Ratio) pour l'extraction des règles positives et négatives d'un contexte de la fouille de données. Ils ont insisté surtout sur le fait que cette mesure permet l'extraction des règles négatives de la forme $X \rightarrow \overline{Y}$, $\overline{X} \rightarrow Y$ et $\overline{X} \rightarrow \overline{Y}$. Notons que eu égard à ses propriétés mathématiques (mesure implicative, orientée et normalisée) la mesure $\mathrm{M_{GK}}$ est parfois appelée ION (Implicative Orientée Normalisée) [TRD04, TR05]. Nous consacrons ce chapitre aux études des propriétés de la mesure $\mathrm{M_{GK}}$ afin de justifier notre intérêt à son utilisation dans la suite de ce travail. Ce chapitre est organisé de la façon suivante. Section 5.2 présente la construction et la définition de la mesure $\mathrm{M_{GK}}$. Nous étudions les principales

propriétés de M_{GK} dans la Section 5.3. Nous faisons des études comparatives de M_{GK} avec d'autres mesures de qualité dans la Section 5.4 avant de conclure dans la Section 5.5.

5.2 Construction et définition

Soient $\mathbb{K} = (\mathcal{E}, \mathcal{A}, \mathcal{R})$ un contexte de la fouille de données, X et Y deux motifs de $\mathbb{K}$. Rappelons les définitions suivantes.

- Pour un motif X, son dual X' est défini par $X' = \{e \in \mathcal{E} | \forall x \in X, x\mathcal{R}e\}$.
- On définit la probabilité p sur l'espace probabilisable discret fini $(\mathcal{E}, \mathcal{P}(\mathcal{E}))$ par, pour tout événement E de $\mathcal{P}(\mathcal{E})$: $p(E) = \frac{|E|}{|\mathcal{E}|}$, où $|A|$ désigne la cardinalité de l'ensemble A.

On dit que :

- X et Y sont incompatibles, si $X' \cap Y' = \emptyset$ soit $p(X' \cap Y') = 0$;
- X favorise Y (équivalent à Y favorise X), si $p(Y'|X') > p(Y')$, où $p(Y'|X')$ désigne la probabilité conditionnelle de Y' sachant X' ; auquel cas X apporterait une information positive au sujet de la réalisation de Y.
- X et Y sont statistiquement indépendants, si $p(Y'|X') = p(Y')$; auquel cas X n'apporterait aucune information sur la réalisation de Y.
- X défavorise Y (équivalent à Y défavorise X), si $p(Y'|X') < p(Y')$; auquel cas X apporterait une information négative au sujet de réalisation de Y.
- X implique logiquement Y, si $X' \subseteq Y'$, *i.e.* $p(Y'|X') = 1$.

Maintenant, nous présentons la construction de la mesure M_{GK}.

(1) Si X défavorise Y, on a : $p(Y'|X') < p(Y')$ soit $p(Y'|X') - p(Y') < 0$. Dans le cas de l'indépendance entre X et Y, on a :

$p(Y'|X') - p(Y') = 0$. Par ailleurs, $0 \leq p(Y'|X')$ implique $-p(Y') \leq p(Y'|X') - p(Y')$. L'égalité est atteinte au cas de l'incompatibilité entre

X et Y. En considérant l'indépendance comme situation limite de dépendance négative, on peut écrire :

$$-p(X) \leq p(Y'|X') - p(Y') \leq 0, \text{ si } X \text{ défavorise } Y$$

soit

$$-1 \leq \frac{p(Y'|X') - p(Y')}{p(Y')}, \leq 0 \text{ si } X \text{ défavorise } Y \tag{5.1}$$

(2) Si X favorise Y, on a $p(Y'|X') > p(Y')$, soit $p(Y'|X') - p(Y') > 0$. Dans le cas de l'indépendance entre X et Y, on a $p(Y'|X') - p(Y') = 0$. Par ailleurs, $p(Y'|X') \leq 1$ implique $p(Y'|X') - p(Y') \leq 1 - p(Y')$. L'égalité est atteinte au cas de l'implication logique entre X et Y. En considérant l'indépendance comme situation limite de dépendance positive, on peut écrire :

$$0 \leq p(Y'|X') - p(Y') \leq 1 - p(Y'), \text{ si } X \text{ favorise } Y$$

soit

$$0 \leq \frac{p(Y'|X') - p(Y')}{1 - p(Y')} \leq 1, \text{ si } X \text{ favorise } Y \tag{5.2}$$

A partir de ces deux propriétés, il apparaît logique de poser la définition ci-dessous.

Définition 46: Soit X et Y deux motifs d'un contexte de fouille de données. On définit la mesure M_{GK} par

$$M_{GK}(X \rightarrow Y) = \begin{cases} \frac{p(Y'|X') - p(Y')}{1 - p(Y')}, & \text{si } X \text{ favorise } Y \\ \frac{p(Y'|X') - p(Y')}{p(Y')}, & \text{si } X \text{ défavorise } Y. \end{cases} \tag{5.3}$$

Notons que cette mesure a été définie indépendamment par Guillaume [Gui00] et Wu et al.[WZZ04].

L'expression de la mesure de qualité M_{GK} en fonction de la mesure Confiance est donnée par :

$$M_{GK}(X \rightarrow Y) = \begin{cases} \frac{Conf(X \rightarrow Y) - Supp(Y)}{1 - Supp(Y)}, & \text{si } X \text{ favorise } Y \\ \frac{Conf(X \rightarrow Y) - Supp(Y)}{Supp(Y)}, & \text{si } X \text{ défavorise } Y. \end{cases} \tag{5.4}$$

5.3 Principales propriétés

Nous présentons dans cette section les principales propriétés de la mesure M_{GK}. Pour cela, considérons un contexte de la fouille de données $\mathbb{K} = (\mathcal{E}, \mathcal{A}, \mathcal{R})$. Nous considérons les motifs X et Y de $\mathbb{K}$ tels que $p(X') \neq 0, p(Y') \neq 0, p(X') \neq 1$ et $p(Y') \neq 1$.

Avant d'établir les propriétés de la mesure M_{GK}, nous rappelons les propriétés liant les dépendances positives et négatives entre deux motifs.

Le lemme ci-après montre que la notion de dépendance positive (resp. négative) entre deux motifs est une relation symétrique.

Lemme 2: (1) X favorise Y, si et seulement si Y favorise X.

(2) X défavorise Y, si et seulement si Y défavorise X.

Démonstration :

$$\begin{array}{lll} (1)\ X \text{ favorise } Y & \Leftrightarrow & p(Y'|X') > p(Y') \\ & \Leftrightarrow & p(Y' \cap X') > p(X')p(Y') \\ & \Leftrightarrow & p(X'|Y') > p(X') \\ & \Leftrightarrow & Y \text{ favorise } X. \end{array}$$

(2) se démontre de la même manière. □

Remarque 14: Rappelons qu'en général, malgré la réciprocité de l'éventuelle dépendance statistique entre deux motifs, le degré de dépendance ne sont pas systématiquement égaux. En effet, en général $p(Y'|X') - p(Y') \neq p(X'|Y') - p(X')$, de même $\frac{p(Y'|X')p(Y') - p(Y')}{1-p(Y')} \neq \frac{p(X'|Y') - p(X')}{1-p(X')}$ (cf. Proposition 29).

Le lemme ci-dessous précise les liens entre les notions de dépendance positive et dépendance négative.

Lemme 3: 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.

Démonstration :

(1) Montrons que (i) $\Leftrightarrow$ (ii)

$$\begin{aligned} X \text{ défavorise } Y &\Leftrightarrow p(Y'|X') < p(Y') \\ &\Leftrightarrow -p(Y'|X') > -p(Y') \\ &\Leftrightarrow 1 - p(Y'|X') > 1 - p(Y') \\ &\Leftrightarrow p(\overline{Y}'|X') > p(\overline{Y}') \\ &\Leftrightarrow X \text{ favorise } \overline{Y}. \end{aligned}$$

Montrons maintenant que que (i) $\Leftrightarrow$ (iii)

$$\begin{aligned} X \text{ défavorise } Y &\Leftrightarrow Y \text{ défavorise } X \\ &\Leftrightarrow p(X'|Y') < p(X') \\ &\Leftrightarrow -p(X'|Y') > -p(X') \\ &\Leftrightarrow 1 - p(X'|Y') > 1 - p(X') \\ &\Leftrightarrow p(\overline{X}'|Y') > p(\overline{X}') \\ &\Leftrightarrow Y \text{ favorise } \overline{X} \\ &\Leftrightarrow \overline{X} \text{ favorise } Y. \end{aligned}$$

(2) Montrons que (i) $\Leftrightarrow$ (ii)
Il suffit d'appliquer (1) à X défavorise $\overline{Y}$.
En effet, X défavorise $\overline{Y} \Leftrightarrow X$ favorise $\overline{\overline{Y}}$, donc X favorise Y.
Les autres équivalences se démontrent de la même manière. □

Les résultats de la proposition suivante résultent de la définition de M_{GK}.

Proposition 26: (**Situations de référence**) Pour tous motifs X et Y, on a :

- X et Y sont incompatibles, si et seulement si $M_{GK}(X \rightarrow Y) = -1$;
- X défavorise Y, si et seulement si $-1 < M_{GK}(X \rightarrow Y) < 0$;
- X et Y sont indépendants, si et seulement si $M_{GK}(X \rightarrow Y) = 0$;
- X favorise Y, si et seulement si $0 < M_{GK}(X \rightarrow Y) < 1$;
- X implique logiquement Y, si et seulement si $M_{GK}(X \rightarrow Y) = 1$.

Les propriétés ci-dessus expriment le fait que M_{GK} prend ses valeurs sur l'intervalle $[-1, 1]$ tout en reflétant les situations de reférence 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.

Elle peut considérer à la fois comme une mesure de l'écart à l'indépendance et de degré d'implication (dépendance orientée) entre la prémisse et le conséquent de la règle. À part les cinq situations de référence mentionnées ci-dessus, Blanchard et al. [BGBG05] considèrent une autre situation de référence à savoir la situation d'équilibre ou d'incertitude maximale (*i.e.*, $|X' \cap Y'| = |X' \cap \overline{Y'}|$).

Une mesure de qualité est dite mesure de *déviation d'équilibre* si elle prend une valeur constante quand le nombre d'exemples et de contre-exemples de la règle sont égaux [BGBG05]. La Proposition 27 ci-dessous montre que la mesure de qualité M_{GK} vérifie la situation de référence à l'équilibre.

Proposition 27: (**Situation de référence à l'équilibre**) [TRD05, DRT07]

$$\text{À l'équilibre : } M_{GK}(X \rightarrow Y) \approx \frac{1}{2}$$

On peut assimiler que la mesure M_{GK} est une mesure de déviation d'équilibre pour n suffisamment grand, *i.e.*, pour un grand volume de données.

Démonstration : Notons qu'une règle pourrait être intéressante si la prémisse favorise le conséquent. Ainsi, considérons une règle d'association $X \rightarrow Y$ telle que X favorise Y. À l'équilibre, $|X' \cap Y'| = |X' \cap \overline{Y}'|$ et notons $n = |\mathcal{E}|$. On a donc :

$$\begin{aligned} M_{GK}(X \rightarrow Y) &= \frac{p(Y'|X') - p(Y')}{1 - p(Y')} \\ &= \frac{\frac{1}{2} - \frac{|Y'|}{n}}{1 - \frac{|Y'|}{n}} \end{aligned}$$

Dans la plupart de cas n est suffisamment grand, donc nous pouvons faire l'approximation :

$$\begin{aligned} M_{GK}(X \rightarrow Y) &= \frac{1}{2} - \frac{|Y'|}{n} 1 - \frac{|Y'|}{n} \\ &= \frac{1}{2} - \frac{|Y'|}{2n} + \frac{|Y'|}{n^2} \\ &= \frac{1}{2} + o(\frac{1}{n}) \end{aligned}$$

Ce qui démontre les résultats. □

La proposition suivante est facile à démontrer.

Proposition 28: Le mesure M_{GK} est sensible à la taille de données, donc elle est une mesure statistique.

La Proposition 29 ci-dessous montre que la mesure de qualité M_{GK} n'est pas une mesure symétrique.

Proposition 29: (i) Si X favorise Y, nous avons la relation :

$$M_{GK}(Y \rightarrow X) = \frac{1 - p(Y')}{1 - p(X')} \frac{p(X')}{p(Y')} M_{GK}(X \rightarrow Y). \tag{5.5}$$

(ii) Si X défavorise Y, nous avons la relation :

$$M_{GK}(Y \rightarrow X) = M_{GK}(X \rightarrow Y) \tag{5.6}$$

Démonstration : (i) Si X favorise Y, nous avons

$$\begin{array}{rcl} M_{GK}(Y \rightarrow X) & = & \frac{p(X'|Y') - p(X')}{1 - p(X')} \\ & = & \frac{p(X' \cap Y') - p(X')p(Y')}{p(Y')(1 - p(X'))} \\ & = & \frac{p(X')}{p(Y')} \frac{1 - p(Y')}{1 - p(X')} \frac{p(X' \cap Y') - p(X')p(Y')}{p(X')(1 - p(Y'))} \\ & = & \frac{p(X')}{p(Y')} \frac{1 - p(Y')}{1 - p(X')} M_{GK}(X \rightarrow Y) \end{array}$$

(ii) Si X défavorise Y, nous avons

$$\begin{array}{rcl} M_{GK}(Y \rightarrow X) & = & \frac{p(X'|Y') - p(X')}{p(X')} \\ & = & \frac{p(X' \cap Y') - p(X')p(Y')}{p(Y')(p(X')} \\ & = & M_{GK}(X \rightarrow Y) \end{array}$$

Ce qui démontre les résultats. □

Corollaire 7: Du point (i) de la Proposition 29, la mesure de qualité M_{GK} est non symétrique.

Proposition 30: (i) Si X favorise Y, nous avons la relation d'équivalence des deux règles contraposées :

$$M_{GK}(\overline{Y} \rightarrow \overline{X}) = M_{GK}(X \rightarrow Y). \tag{5.7}$$

(ii) Si X défavorise Y, nous avons la relation :

$$M_{GK}(\overline{Y} \rightarrow \overline{X}) = \frac{p(X')p(Y')}{(1 - p(X'))(1 - p(Y'))} M_{GK}(X \rightarrow Y) \tag{5.8}$$

Démonstration : (i) Si X favorise Y, nous avons

$$\begin{aligned}
\mathrm{M_{GK}}(\overline{Y}\rightarrow\overline{X}) &= \frac{p(\overline{X}'|\overline{Y}')-p(\overline{X}')}{1-p(\overline{X}')} \\
&= \frac{1-p(X'|\overline{Y}')-1+p(X')}{p(X')} \\
&= \frac{-p(X'\cap\overline{Y}')+p(X')p(\overline{Y}')}{p(X')(1-p(Y'))} \\
&= \frac{-p(X')+p(X'\cap Y')+p(X')-p(X')p(Y')}{p(X')(1-p(Y'))} \\
&= \frac{p(X'\cap Y')-p(X')p(Y')}{p(X')(1-p(Y'))} \\
&= \mathrm{M_{GK}}(X\rightarrow Y)
\end{aligned}$$

(ii) Si X défavorise Y, nous avons

$$\begin{aligned}
\mathrm{M_{GK}}(\overline{Y}\rightarrow\overline{X}) &= \frac{p(\overline{X}'|\overline{Y}')-p(\overline{X}')}{p(\overline{X}')} \\
&= \frac{1-p(X'|\overline{Y}')-1+p(X')}{1-p(X')} \\
&= \frac{-p(X'\cap\overline{Y}')+p(X')p(\overline{Y}')}{(1-p(X'))(1-p(Y'))} \\
&= \frac{-p(X')+p(X'\cap Y')+p(X')-p(X')p(Y')}{(1-p(X'))(1-p(Y'))} \\
&= \frac{p(X'\cap Y')-p(X')p(Y')}{(1-p(X'))(1-p(Y'))} \\
&= \frac{p(X')p(Y')}{(1-p(X'))(1-p(Y'))}\frac{p(X'\cap Y')-p(X')p(Y')}{p(X')p(Y')} \\
&= \frac{p(X')p(Y')}{(1-p(X'))(1-p(Y'))}\mathrm{M_{GK}}(X\rightarrow Y)
\end{aligned}$$

Ce qui démontre les résultats. □

Le corollaire ci-dessous est une conséquence de l'équation (5.7).

Corollaire 8: $\mathrm{M_{GK}}$ est une mesure implicative dans la zone d'attraction.

La mesure $\mathrm{M_{GK}}$ s'avère ainsi fort intéressante pour mesurer la dépendance positive entre deux motifs : soit entre motifs positifs, soit entre motifs positif et négatif. En effet, par le Lemme 3, on peut toujours se ramener au cas où l'un des des motifs (prémisse ou conséquent) favorise l'autre. Donc, dans la pratique, nous retenons tout simplement que $\mathrm{M_{GK}}$ est implicative.

Proposition 31: Soient X et Y deux motifs positifs. On a l'égalité suivante :

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

Démonstration : Raisonnons par distinction des deux cas : (1) X favorise Y et (2) X défavorise Y.

(1) Si X favorise Y (donc X défavorise Y), on a :

$$\begin{aligned}\mathrm{M_{GK}}(X{\rightarrow}\overline{Y}) &= \frac{p(\overline{Y}'|X')-p(\overline{Y}')}{p(\overline{Y}')}\\ &= \frac{1-p(Y'|X')-1+p(Y')}{1-p(Y')}\\ &= \frac{p(Y')-p(Y'|X')}{1-p(Y')}\\ &= -\mathrm{M_{GK}}(X{\rightarrow}Y)\end{aligned}$$

(2) Si X défavorise Y (donc X favorise $\overline{Y}$), on a :

$$\begin{aligned}\mathrm{M_{GK}}(X{\rightarrow}\overline{Y}) &= \frac{p(\overline{Y}'|X')-p(\overline{Y}')}{1-p(\overline{Y}')}\\ &= \frac{1-p(Y'|X')-1+p(Y')}{1-p(Y')}\\ &= \frac{p(Y')-p(Y'|X')}{p(Y')}\\ &= -\mathrm{M_{GK}}(X{\rightarrow}Y)\end{aligned}$$

Ce qui démontre les résultats. □

Le corollaire ci-dessous est une conséquence de la proposition précédente. La mesure $\mathrm{M_{GK}}$ permet de calculer les règles négatives valides à partir de règles positives correspondantes.

Corollaire 9: Soient $\alpha \in [0,1]$, X et Y deux motifs positifs tels que X défavorise Y. On a les résultats suivants

$$-1 < \mathrm{M_{GK}}(X{\rightarrow}Y) \leq \alpha \Leftrightarrow \alpha \leq \mathrm{M_{GK}}(X{\rightarrow}\overline{Y}) < 1.$$

Les relations entre les règles négatives à droite et les règles négatives à gauche sont données par la proposition suivante.

Proposition 32: (1) Si X défavorise Y (donc, X favorise $\overline{Y}$ et aussi $\overline{X}$ favorise Y), on a :

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

(2) Si X favorise Y (donc, X défavorise $\overline{Y}$ et aussi $\overline{X}$ défavorise Y), on a :

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

Démonstration : Nous commençons par démontrer le point (1) de cette proposition.

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

Nous montrons maintenant le point (2) de cette proposition.

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

Ce qui démontre les résultats. □

Les liens entre les valeurs de $\mathrm{M_{GK}}$ pour les règles d'association qui se trouvent sur une chaîne du le treillis de motifs sont données par la proposition suivante.

Proposition 33: (multiplicativité)

(i) Si $X\subseteq Y\subseteq Z$ alors

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

(ii) Si $X_1\subseteq X_2\subseteq ...\subseteq X_i\subseteq X_{i+1}\subseteq ...\subseteq X_p$ alors

$$\mathrm{M_{GK}}(X_1\rightarrow X_p) = \prod_{i=1}^{p-1}\mathrm{M_{GK}}(X_i\rightarrow X_{i+1})$$

$\mathrm{M_{GK}}$ est multiplicative sur une chaîne.

Démonstration : Il suffit de démontrer le point (i). Le point (ii) se généralise très facilement. Les inclusions $X\subseteq Y\subseteq Z$ impliquent $Z'\subseteq Y'\subseteq X'$.

Donc, $p(Y'|X') = \frac{p(X'\cap Y')}{p(X')} = \frac{p(Y')}{p(X')} \geq p(Y')$.
Les trois motifs X, Y et Z se favorisent deux à deux. Par ailleurs, comme ces trois motifs se favorisent deux à deux, on a :

$$\begin{aligned} \mathrm{M_{GK}}(X\rightarrow Y) &= \frac{\frac{p(X'\cap Y')}{p(X')} - p(Y')}{1-p(Y')} \\ &= \frac{\frac{p(Y')}{p(X')} - p(Y')}{1-p(Y')} \\ &= \frac{p(Y')(1-p(X')}{p(X')(1-p(Y'))}. \end{aligned}$$

Nous avons donc,

$$\begin{aligned} \mathrm{M_{GK}}(X\rightarrow Y)\mathrm{M_{GK}}(Y\rightarrow Z) &= \frac{p(Y')(1-p(X')}{p(X')(1-p(Y'))}\frac{p(Z')(1-p(Y')}{p(Y')(1-p(Z'))} \\ &= \frac{p(Z')(1-p(X')}{p(X')(1-p(Z'))} \\ &= \mathrm{M_{GK}}(X\rightarrow Z). \end{aligned}$$

Ce qui démontre le résultat. □

Remarque 15: Grâce à cette propriété multiplicative, l'utilisation de la mesure de qualité $\mathrm{M_{GK}}$ permet l'élaguer des branches de transitivité sur un diagramme de Hasse partiel ou dans un graphe sous-treillis a priori dense. Ce qui permet d'améliorer la lisibilité d'un graphe implicatif.

Le corollaire suivant est une conséquence de la Proposition 33 ci-dessus.

Corollaire 10: Soient $X_1, X_2, ..., X_i, X_{i+1}, ..., X_p$ des motifs tels que $X_1 \subseteq X_2 \subseteq ... \subseteq X_i \subseteq X_{i+1} \subseteq ... \subseteq X_p$.

(i) Si $X_1 \rightarrow X_p$ est $(\mathrm{M_{GK}}, \alpha)$-valide alors $\forall i, j \in \{1, ..., p\}$ avec $i < j$, $X_i \rightarrow X_j$ est $(\mathrm{M_{GK}}, \alpha)$-valide.

(ii) Si il existe $i, j \in \{1, ..., p\}$ tels que $X_i \rightarrow X_j$ est non $(\mathrm{M_{GK}}, \alpha)$-valide alors $\forall l, k \in \{1, ..., p\}$ tels que $l \leq i$ et $j \leq k$, $X_l \rightarrow X_k$ est aussi non $(\mathrm{M_{GK}}, \alpha)$-valide.

La proposition ci-dessous montre que la mesure de qualité $\mathrm{M_{GK}}$ permet de sélectionner moins de règles que la mesure Confiance pour un même seuil de validité.

Proposition 34: Soient X et Y deux motifs d'un contexte de la fouille de données. On a l'inégalité suivante :

$$M_{GK}(X \rightarrow Y) \leq Conf(X \rightarrow Y).$$

Démonstration : Il suffit de démontrer le cas où X favorise Y (car $M_{GK}(X \rightarrow Y) \leq 0$ si X défavorise Y alors que $Conf(X \rightarrow Y) \in [0, 1]$).

$$\begin{aligned} M_{GK}(X \rightarrow Y) - Conf(X \rightarrow Y) &= \frac{\frac{p(Y'|X') - p(Y')}{1-p(Y')} - p(Y'|X')}{} \\ &= \frac{\frac{p(Y'|X') - p(Y') - p(Y'|X') + p(Y'|X')p(Y')}{1-p(Y')}}{} \\ &= \frac{\frac{p(Y')(-1 + p(Y'|X'))}{1-p(Y')}}{} \\ &\leq 0 \end{aligned}$$

Ce qui démontre le résultat. □

5.4 Comparaison à d'autres mesures de qualité

Nous présentons dans la présente section des études comparatives de la mesure de qualité M_{GK} avec d'autres mesures de qualité qu'on peut trouver dans la littérature :

- Confiance : c'est la mesure souvent utilisée par les différentes méthodes de la fouille des règles d'association
- ϕ-coefficient : c'est la mesure de qualité la plus utilisée par les statisticiens ;
- Loevinger : S. Guillaume définit M_{GK} à partir de la mesure de Loevinger, par ailleurs, cette mesure est la plus ancienne des mesures de qualité proposées dans la littérature ;
- Lift : c'est une mesure qui a été construite en vue de remédier les faiblesses de la mesure de qualité Confiance ;
- Conviction : c'est une mesure implicative comme M_{GK} mais elle n'est pas bornée.

5.4.1 M_{GK} et Confiance

Le Tableau 5.1 compare les comportements de mesures M_{GK} et Confiance. La mesure Confiance tient compte des situations de référence au point d'incompatibilité et au point d'implication logique entre la prémisse et le conséquent de la règle. Comme la valeur de la Confiance au point d'indépendance est égale $p(Y')$ (donc variable), elle ne permet pas de nous renseigner, si nous sommes dans la zone d'attraction ou de répulsion, ce qui n'est pas le cas pour M_{GK}. C'est la raison pour laquelle les processus d'extraction des règles utilisant la Confiance peut sélectionner des règles qui n'ont pas d'intérêt (le cas où la prémisse et le conséquent se défavorisant ou ils sont indépendants) comme soulevé dans [BMUT97, Gui00, LT04]. En ce sens, M_{GK} permet de remédier à cette faiblesse de la Confiance. Par ailleurs, pour toute règle d'association $M_{GK}(X \rightarrow Y) \leq \text{Conf}(X \rightarrow Y)$.

En outre, pour un même seuil de validité, la mesure de qualité M_{GK} sélectionne moins de règles que la mesure de qualité Confiance seule.

Mesure	Incompati.	Répulsion	Indép.	Attraction	Implication
M_{GK}	-1	négative	0	positive	1
Confiance	0	positive	$p(Y')$	positive	1

Tab. 5.1: *Comparaison de valeurs prises par les mesures* M_{GK} *et Confiance*

5.4.2 M_{GK} et mesure de Loevinger

Rappelons que la mesure de Loevinger [Loe47] est définie par

$$Loevinger(X \rightarrow Y) = \frac{p(Y'|X') - p(Y')}{1 - p(Y)'}.$$

Ces mesures de qualité sont identiques dans le cas où la prémisse et le conséquent d'une règle se favorisent et elles sont différentes quand la prémisse et le conséquent d'une règle d'association se défavorisent. Contrairement à M_{GK}, la mesure de Loevinger n'est pas constante en cas d'incompatibilité. Même si Loevinger permet de nous renseigner si nous sommes dans la zone d'attraction ou de répulsion, elle ne permet pas de prédire sur le degré de répulsion entre la prémisse et le conséquent (cf. Tableau 5.2) d'une règle d'association.

Mesure	Incompati.	Répulsion	Indép.	Attraction	Implica.
M_{GK}	-1	négative	0	positive	1
Loevinger	$-\frac{p(Y')}{1-p(Y')}$	négative	0	positive	1

Tab. 5.2: *Comparaison de valeurs prises par les mesures* M_{GK} *et Loevinger*

5.4.3 M_{GK} et Lift

La mesure Lift [BMUT97] est définie par

$$Lift(X \rightarrow Y) = \frac{p(X' \cap Y')}{p(X')p(Y')}.$$

Nous présentons dans un même tableau les valeurs prises par ces deux mesures dans les zones et les points de référence (cf. Tableau 5.3). M_{GK} et Lift permettent de nous renseigner si nous sommes dans le cas d'incomptibilité, la zone de répulsion, le cas d'indépendance. Contrairement à M_{GK}, Lift ne nous permet pas de distinguer l'attraction et l'implication logique entre la prémisse et le conséquent d'une règle d'association. Par ailleurs, Lift est une mesure symétrique. Elle est une mesure de co-occurrence de la prémisse et du conséquent de la règle. Lift ne permet pas de faire le choix entre $X \rightarrow Y$ et $Y \rightarrow X$; au sens de Lift les règles $X \rightarrow Y$ et $Y \rightarrow X$ sont équivalentes. Contrairement à la mesure de qualité M_{GK}, Lift ne rend pas compte de la sémantique de l'implication "si ... alors".

Mesure	Incompati.	Répulsion	Indép.	Attraction	Implica.
M_{GK}	-1	négative	0	positive	1
Lift	0	positive	1	positive	$\frac{1}{p(Y')}$

Tab. 5.3: *Comparaison de valeurs prises par les mesures* M_{GK} *et Lift*

5.4.4 M_{GK} et ϕ-coefficient

La mesure ϕ-coefficient [LGR81] est définie par

$$\phi(X \rightarrow Y) = \frac{p(X' \cap Y') - p(X')p(Y')}{\sqrt{p(X')p(Y')p(\overline{X}')p(\overline{Y}')}}.$$

Le Tableau 5.4 compare les valeurs prises par ces deux mesures de qualité dans des situations de référence.

Mesure	Incompati.	Répulsion	Indép.	Attraction	Implica.
M_{GK}	-1	négative	0	positive	1
ϕ-coefficient	$\sqrt{\frac{p(X')p(Y')}{p(\overline{X}')p(\overline{Y}')}}$	négative	0	positive	$\sqrt{\frac{p(X')p(\overline{Y}')}{p(\overline{X}')p(Y')}}$

Tab. 5.4: *Comparaison de valeurs prises par les mesures* M_{GK} *et* ϕ*-coefficient*

On constate dans le Tableau 5.4 que M_{GK} et ϕ-coefficient nous permettent de nous renseigner si nous sommes dans la zone d'attraction ou dans la zone de répulsion ou au point d'indépendance. Ces deux mesures ont toujours le même signe. Toutefois, comme les valeurs extrêmes (incompatibilité et implication logique) de la mesure ϕ-coefficient sont variables, elle ne permet pas de nous renseigner si l'attraction ou la répulsion entre la prémisse et le conséquent de la règle soit forte ou faible. Par ailleurs, ϕ-coefficient est une mesure symétrique. Elle ne permet pas de distinguer les règles $X \rightarrow Y$ et $Y \rightarrow X$. Ce qui n'est pas le cas pour la mesure M_{GK}. Il semble que M_{GK} présente beaucoup plus d'intérêt par rapport à la mesure ϕ-coefficient.

5.4.5 M_{GK} et Conviction

La mesure Conviction [BMUT97] est définie préferentiellement par

$$Conviction(X \rightarrow Y) = \frac{p(X')p(\overline{Y}')}{p(X' \cap \overline{Y}')}.$$

Nous présentons dans un même tableau les valeurs prises par ces deux mesures dans les zones et les points de référence (cf. Tableau 5.5) La mesure Conviction prend la valeur 1 en cas d'indépendance et elle permet de situer si nous sommes dans la zone de répulsion ou dans la zone d'attraction. Par ailleurs, comme mentionné dans [BMUT97], la Conviction est favorablement une mesure implicative. Elle tient compte de la sémantique d'implication "si ... alors". Toutefois, à défaut d'être non bornée, le comportement de la Conviction en cas d'implication logique (Conviction $=\infty$) entre la prémisse et le conséquent de la règle, elle ne permet pas de nous renseigner si l'attraction est faible ou forte. Ce qui n'est pas le cas pour la mesure implicative M_{GK}.

Mesure	Incompati.	Répulsion	Indép.	Attraction	Implica.
M_{GK}	-1	négative	0	positive	1
Conviction	$p(\overline{Y}')$	positive	1	positive	$+\infty$

Tab. 5.5: *Comparaison des valeurs prises par les mesures* M_{GK} *et Conviction*

5.5 Conclusion

Nous avons présenté la construction et la définition de la mesure M_{GK} introduite indépendamment par Guillaume [Gui00] et Wu et al. [WZZ04]. Nous avons étudié les principales propriétés de cette mesure. Ces études permettent d'apprécier l'intérêt de la mesure M_{GK}. Elle vérifie plusieurs propriétés parmi celles qui sont souhaitées par les chercheurs travaillant sur les études de mesures de qualité des règles d'association. Citons par exemple, M_{GK} est une mesure non symétrique, vérifie les principes de Piatetsky-Shapiro, le principe de Freitas etc. La propriété implicative de M_{GK} permet de sélectionner des règles implicatives. La mesure M_{GK} est une mesure d'écart à l'indépendance et de degré d'implication entre la prémisse et le conséquent d'une règle d'association.

Les comparaisons de la mesure de qualité M_{GK} avec d'autres mesures de qualité que nous avons entreprises dans la présent chapitre nous permettent d'apprécier davantage M_{GK} par rapport à ces mesures. Rappelons que le problème de choix de la (des) mesure(s) de qualité est un problème d'actualité en fouille des règles d'association. Par ailleurs, le caractère implicatif d'une mesure de qualité s'avère souvent souhaité [BMUT97]. Il semble que l'utilisation de la mesure M_{GK} permettrait de sélectionner à la fois des règles positives et négatives intéressantes dans un contexte de la fouille de données.

6. NORMALISATION DE MESURES DE QUALITÉ

6.1 Introduction

Plusieurs algorithmes de la fouille des règles d'association sont disponibles dans la littérature APRIORI [AIS93, AS94], CLOSE [PBTL99a], DIC[BMUT97], CLOSET [PHM00], CHARM [ZH99]. Les mesures de qualité servent à évaluer, à classifier les règles d'association d'un contexte de la fouille de données. Eu égard à la littérature sur la fouille des règles d'association, il s'avère que les mesures Support et Confiance sont les plus utilisées par les différentes méthodes de la fouille des règles d'association. Toutefois, depuis ces dernières années, l'utilisation de ces mesures suscite plusieurs critiques. En effet, d'une part, ces mesures peuvent sélectionner certaines règles sans intérêt (cas de l'indépendance entre la prémisse et le conséquent d'une règle si elle a une valeur de Confiance dépasse le seuil minimum Confiance) [LT04, BMUT97], d'autre part, la mesure Support, considérée comme moteur de processus d'extraction, écarte les règles ayant une valeur faible de Support alors que certaines peuvent avoir une très forte Confiance et présenter un réel intérêt : les petites connaissances [Azé03a]. Pour tenter de pallier cela, plusieurs mesures de qualité ont été proposées. Ce qui engendre de nouveaux problèmes, entre autres, le choix de mesure(s) utilisée(s) pour l'extraction des règles d'un contexte de la fouille de données. Il se pose ainsi naturellement la question suivante "quelle mesure de qualité doit être utilisée pour sélectionner les règles intéressantes d'un contexte de la fouille de données?" Plusieurs critères ont été suggerés pour concevoir et pour choisir une mesure de qualité à utiliser afin de capturer des règles intéressantes. Cependant, il est très difficile de trouver une mesure vérifiant l'ensemble de ces critères.

Dans le présent chapitre, nous proposons des études formelles sur les

mesures de qualité des règles que nous appellerons *normalisation* en vue d'apporter un nouvel éclairage sur les mesures de qualité de règles. Le but est alors de présenter une vue unificatrice des différentes mesures de qualité qui est une poursuite des travaux présentés dans [Tot03]. Le reste de ce chapitre est organisé de la façon suivante. Dans la Section 6.2, nous présentons les motivations et définissons une mesure normalisée. La normalisation de mesures de qualité de règles d'association est présentée dans la Section 6.3. Nous donnons quelques exemples de normalisation de mesures de qualité des règles dans la Section 6.4. La classification des mesures de qualité des règles est présentée dans la Section 6.5 avant de conclure dans la Section 6.6.

6.2 Motivation et Définitions

"Normer", "centrer" une mesure de qualité des règles sont des termes utilisés vaguement par les chercheurs travaillant dans le domaine de la fouille des règles d'association. Pour corriger les faiblesses de la mesure Confiance, Lallich et Teytaud [LT04] définissaient la mesure Confiance centrée en enlevant de la Confiance la probabilité du conséquent de la règle pour avoir une référence en cas d'indépendance entre la prémisse et le conséquent de la règle. La mesure de Loevinger [Loe47], l'une des plus anciennes mesures de qualité répertoriées dans le domaine de fouille de données, normalise la Confiance centrée. Elle permet de pallier à un des défauts de la Confiance. Brin et al. [BMUT97] préconisaient la conviction qui est une mesure implicative normlisée. Cependant, il s'avère que la normalisation évoquée et souhaitée n'est pas explicitée. Toutefois, comme la Conviction prend la valeur limite ∞ en cas d'implication logique entre la prémisse et le conséquent d'une règle, elle ne permet pas d'indiquer à partir de quelles valeurs de la Conviction "une règle est dite convaincante". En fait, les mesures de qualité ont des comportements hétérogènes face aux critères souhaités pour une bonne mesure de qualité des règles. On peut observer d'importantes variations entre les formules et des grandes différences dans les ensembles des valeurs prises par une mesure. Pour mettre en lumière cela, considérons le contexte de fouille de données présenté dans le Tableau 6.1 formé de cinq attributs et six entités.

Le Tableau 6.2 présente les valeurs prises par quelques mesures de qualité pour certaines règles d'association considérées. On constate que les valeurs prises par ces mesures de qualité sont distribuées entre -1 et $+\infty$. De plus,

	A	B	C	D	E
e_1	1	1	1	1	0
e_2	0	1	1	0	0
e_3	1	0	1	1	1
e_4	1	1	1	0	1
e_5	0	0	0	1	1
e_6	1	0	0	1	1

Tab. 6.1: *Contexte binaire*

Règle	Support $[0;1]$	Confiance $[0;1]$	M_{GK} $[-1;1]$	Conviction $[0;+\infty]$	Jaccard $[0,1]$
$BC \rightarrow DE$	0	0	-1	$\frac{1}{2}$	0
$DE \rightarrow A$	$\frac{1}{3}$	$\frac{1}{2}$	$-\frac{1}{4}$	$\frac{1}{2}$	$\frac{2}{5}$
$BC \rightarrow ACD$	$\frac{1}{6}$	$\frac{1}{3}$	0	1	$\frac{1}{4}$
$ACD \rightarrow ABC$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{4}{3}$	$\frac{1}{3}$
$ACD \rightarrow A$	$\frac{1}{3}$	1	1	$+\infty$	$\frac{1}{2}$

Tab. 6.2: *Valeurs prises par quelques mesures de qualité*

Situation de référence	Support $[0;1]$	Confiance $[0;1]$	M_{GK} $[-1;1]$	Conviction $[0;+\infty]$	Jaccard $[0,1]$
Incompati.	0	0	-1	$p(\overline{Y}')$	0
Répulsion	positive	positive	négative	positive	positive
Indép.	$p(X')p(Y')$	$p(Y')$	0	1	$\frac{p(X')p(Y')}{p(X')+p(Y')-p(X')p(Y')}$
Attraction	positive	positive	positive	positive	positive
Implication	$p(X')$	1	1	$+\infty$	$\frac{p(X')}{p(Y')}$

Tab. 6.3: *Comportements de quelques mesures de qualité*

certaines mesures de qualité prennent des valeurs positives indépendament du fait que la prémisse favorise le conséquent. En fait, nous avons le Tableau 6.3 qui présente les comportements de ces différentes mesures de qualité dans les situations de référence. Prenons par exemple le cas de la Confiance : elle n'a pas une valeur fixe en cas d'indépendance entre la prémisse et le conséquent d'une règle. Ce qui entraîne la possibilité de sélectionner des règles où la prémisse et le conséquent sont indépendants et même si la prémisse défavorise le conséquent pourvu qu'elles vérifient les conditions de Support et Confiance, alors que les règles de ce type n'ont aucun intérêt pour l'utilisateur. Certaines mesures ne prennent pas de valeurs fixes à l'implication, ce qui engendre la difficulté de définir un seuil minimum. D'où la possibilité d'écarter certaines règles intéressantes (en cas de l'implication entre la prémisse et le conséquent). L'objectif de la normalisation est alors de ramener les valeurs d'une mesure de qualité sur l'intervalle $[-1, 1]$ tout en reflétant 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. Avant de donner la définition proposée pour une mesure normalisée, nous rappelons les définitions concernant les situations de référence d'une règle en termes de probabilités.

Définition 47: Soient X et Y des motifs d'un contexte de la fouille de données. On dit que :

(i) X et Y sont incompatibles si et seulement si $X' \cap Y' = \emptyset$ (donc $p(Y'|X') = 0$) ;

(ii) X et Y sont négativement dépendants ou X et Y se défavorisent mutuellement si et seulement si $p(Y'|X') < p(Y')$ (ce qui est équivalent à $p(X'|Y') < p(X')$) ;

(iii) X et Y sont indépendants si et seulement si $p(Y'|X') = p(Y')$;

(iv) X et Y sont positivement dépendants ou X et Y se favorisent mutuellement si et seulement si $p(Y'|X') > p(Y')$ (ce qui est équivalent à $p(X'|Y') > p(X')$) ;

(v) X implique logiquement Y si et seulement si $X' \subseteq Y'$ (donc $p(Y'|X') = 1$).

Eu égard aux objectifs de la normalisation cités ci-dessus, nous posons la définition d'une mesure de qualité normalisée de la façon suivante.

Définition 48: [DRT07] Soit $X \rightarrow Y$ une règle d'association. Une mesure de qualité μ est dite *normalisée* si elle vérifie les cinq conditions ci-dessous :

(i) $\mu(X \rightarrow Y) = -1$ si et seulement si X et Y sont incompatibles ;

(ii) $-1 < \mu(X \rightarrow Y) < 0$ si et seulement si X défavorise Y ou X et Y sont négativement dépendants ;

(iii) $\mu(X \rightarrow Y) = 0$ si et seulement si X et Y sont indépendants ;

(iv) $0 < \mu(X \rightarrow Y) < 1$ si et seulement si X favorise Y ou X et Y sont positivement dépendants ;

(v) $\mu(X \rightarrow Y) = 1$ si et seulement si X implique logiquement Y.

Exemple 14:

- Les deux mesures de qualité suivantes sont des mesures normalisées.
  - La mesure de qualité $\mathrm{M_{GK}}$ [Gui00] définie par :

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

  - La mesure de Zhang [Zha00] définie par :

$$Zhang(X \rightarrow Y) = \frac{p(X' \cap Y') - p(X')p(Y')}{max\{p(X' \cap Y')p(\overline{Y}'); p(Y')p(X' \cap \overline{Y}')\}}$$

- Par contre, la mesure $Lift$ [BMS97] définie par :

$$Lift = \frac{p(Y'|X')}{p(Y')}$$

  est une mesure non normalisée. En effet, par exemple, $Lift(X \rightarrow Y) = 0$ quand X et Y sont incompatibles.

Remarque 16: Notons que les conditions (ii), (iii) et (iv) de la Définition 48 sont les critères de Piatetsky-Shapiro [PS91] pour une bonne mesure de qualité. Nous avons ajouté deux conditions supplémentaires, à savoir la valeur -1 en cas d'incompatibilité et la valeur 1 en cas de l'implication logique entre la prémisse et le conséquent de la règle afin d'encadrer les valeurs prises par une mesure de qualité. Cet encadrement permet d'indiquer si l'attraction ou la répulsion entre la prémisse et le conséquent de la règle est forte ou faible. Par exemple, pour une règle d'association $X \rightarrow Y$, une valeur de mesure de qualité voisine de 1 indique que l'attraction est forte entre la prémisse et le conséquent, donc la règle est intéressante. Par contre, une valeur de mesure voisine de -1 indique que la répulsion est forte entre la prémisse et le conséquent, dans ce cas les règles négatives à droite $X \rightarrow \overline{Y}$ et $Y \rightarrow \overline{X}$ sont intéressantes. Ce qui n'est pas le cas si la mesure n'est pas bornée.

6.3 Caractérisation

Il est évident que ce ne sont pas toutes les mesures disponibles dans la littérature qui sont normalisées. Toutefois, se pose la question de savoir l'existence de moyen pour rendre normalisée une mesure qui ne l'est pas. La présente section répond positivement à cette question. Nous donnons une condition nécessaire et suffisante pour qu'une mesure de qualité soit normalisable. Considérons une mesure de qualité μ, désignons par μ_n la mesure normalisée associée à la mesure μ si elle existe. Pour rendre facile l'interprétation d'une règle, la normalisation de μ consisterait à ramener ses valeurs sur l'intervalle $[-1, 1]$ de telle sorte que la valeur -1 corresponde à l'incompatibilité, les valeurs strictement comprises entre -1 et 0 correspondent à la répulsion ou la dépendance négative, la valeur 0 corresponde à l'indépendance, les valeurs strictement comprises entre 0 et 1 correspondent à l'attraction ou à la dépendance positive orientée et la valeur 1 corresponde à l'implication logique entre la prémisse et le conséquent d'une règle $X \rightarrow Y$. Soit x_f (resp. y_f) le coefficient de multiplication (resp. de centrage) de μ, dans le cas où X favorise Y. De façon similaire, posons x_d (resp. y_d) le coefficient de multiplication (resp. de centrage) dans le cas où X défavorise Y. On a donc :

$$\mu_n(X \rightarrow Y) = \begin{cases} x_f.\mu(X \rightarrow Y) + y_f & si \quad X \quad favorise \quad Y \\ x_d.\mu(X \rightarrow Y) + y_d & si \quad X \quad \text{défavorise} \quad Y \end{cases}$$

Ces quatre coefficients se déterminent par passage aux limites dans des situations de référence (incompatibilité, indépendance et implication logique) du fait de la continuité de l'évolution dans les deux zones : attraction (dépendance positive) et répulsion (dépendance négative). Posons $\mu_{imp}(X \to Y)$ la valeur de $\mu(X \to Y)$ à l'implication, $\mu_{ind}(X \to Y)$ celle de $\mu(X \to Y)$ à l'indépendance et $\mu_{inc}(X \to Y)$ la valeur de $\mu(X \to Y)$ à l'incompatibilité.
Au cas où X favorise Y, on obtient :

$$\begin{cases} x_f \mu_{imp}(X \to Y) + y_f = 1 & \text{implication logique} \\ x_f \mu_{ind}(X \to Y) + y_f = 0 & \text{indépendance à droite} \end{cases}$$

Au cas où X défavorise Y, on obtient :

$$\begin{cases} x_d \mu_{ind}(X \to Y) + y_d = 0 & \text{indépendance à gauche} \\ x_d \mu_{inc}(X \to Y) + y_d = -1 & \text{incompatibilité} \end{cases}$$

Nous pouvons écrire le système d'équations linéaires suivant :

$$\begin{cases} x_f . \mu_{imp}(X \to Y) + y_f = 1 \\ x_f . \mu_{ind}(X \to Y) + y_f = 0 \\ x_d . \mu_{ind}(X \to Y) + y_d = 0 \\ x_d . \mu_{inc}(X \to Y) + y_d = -1 \end{cases} \tag{6.1}$$

L'écriture matricielle de l'équation (6.1) est donnée par l'équation (6.2) :

$$\begin{pmatrix} \mu_{imp}(X \to Y) & 1 & 0 & 0 \\ \mu_{ind}(X \to Y) & 1 & 0 & 0 \\ 0 & 0 & \mu_{ind}(X \to Y) & 1 \\ 0 & 0 & \mu_{inc}(X \to Y) & 1 \end{pmatrix} \begin{pmatrix} x_f \\ y_f \\ x_d \\ y_d \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \end{pmatrix} \tag{6.2}$$

Posons M la matrice associée à ce système. On a donc,

$$M = \begin{pmatrix} \mu_{imp}(X \to Y) & 1 & 0 & 0 \\ \mu_{ind}(X \to Y) & 1 & 0 & 0 \\ 0 & 0 & \mu_{ind}(X \to Y) & 1 \\ 0 & 0 & \mu_{inc}(X \to Y) & 1 \end{pmatrix}$$

Pour que l'équation (6.2) admet une solution unique il faut et il suffit que le déterminant de la matrice M soit fini et non nul. Ainsi, nous avons la caractérisation de mesures de qualité normalisables résultant de l'existence de solution de l'équation (6.2).

Théorème 21: Une mesure de qualité μ est normalisable si et seulement si, pour toute règle $X \rightarrow Y$, les conditions suivantes sont vérifées :
(i) les quantités $\mu_{imp}(X \rightarrow Y)$, $\mu_{ind}(X \rightarrow Y)$ *et* $\mu_{inc}(X \rightarrow Y)$ sont finies ;
(ii) les inégalités suivantes sont vérifiées $\mu_{imp}(X \rightarrow Y) \neq \mu_{ind}(X \rightarrow Y)$, $\mu_{ind}(X \rightarrow Y) \neq \mu_{inc}(X \rightarrow Y)$.

Démonstration : Le déterminant de la matrice M associée à l'équation (6.2) est égal à $(\mu_{imp}(X \rightarrow Y) - \mu_{ind}(X \rightarrow Y))(\mu_{ind}(X \rightarrow Y) - \mu_{inc}(X \rightarrow Y))$. D'où le résultat du théorème. □

Remarque 17:
- Le système d'équation (6.1) ne peut pas avoir une infinité de solutions. En effet, si $det(M) = 0$, pour assurer l'infinité de solution, il faut que le second membre du système soit nul. Ce qui n'est pas le cas.

- Les règles d'association considérées sont des règles $X \rightarrow Y$ telles que $p(X') \neq 0$, $p(Y') \neq 0$, $p(X') \neq 1$ et $p(Y') \neq 1$. En effet, si $p(X') = 1$ donc les attributs qui constituent X sont présents dans toutes les entités, donc le motif X ne porte aucune information nouvelle à l'utilisateur. Par ailleurs, si $p(X') = 0$ la présence simultanée des attributs qui composent X ne se réalise dans aucune entité, donc le motif X ne porte aucune information nouvelle à l'utilisateur.

La proposition suivante établit l'expression des coefficients de transformation pour une mesure de qualité normalisable.

Proposition 35: Soient μ une mesure de qualité normalisable et $X \rightarrow Y$ une règle d'association. Les coefficients de multiplication et de centrage sont donnés par les expressions ci-dessous :

$$x_f = \frac{1}{\mu_{imp}(X \rightarrow Y) - \mu_{ind}(X \rightarrow Y)},\ y_f = -\frac{\mu_{ind}(X \rightarrow Y)}{\mu_{imp}(X \rightarrow Y) - \mu_{ind}(X \rightarrow Y)}\ ;$$
$$x_d = \frac{1}{\mu_{ind}(X \rightarrow Y) - \mu_{inc}(X \rightarrow Y)},\ y_d = -\frac{\mu_{ind}(X \rightarrow Y)}{\mu_{ind}(X \rightarrow Y) - \mu_{inc}(X \rightarrow Y)}.$$

Remarque 18: Il est à noter que les coefficients x_f, x_d, y_f *et* y_d ne dépendent que des probabilités $p(X')$ et $p(Y')$ de la même manière que les quantités $\mu_{imp}(X \rightarrow Y), \mu_{ind}(X \rightarrow Y)$ et $\mu_{inc}(X \rightarrow Y)$.

6.4 Exemples de normalisation de mesures de qualité

Pour illustrer le processus de normalisation des mesures de qualité, voici quelques détails de calcul de la normalisée associée à certaines mesures de qualité. Soit $X \rightarrow Y$ une règle d'association d'un contexte de la fouille de données.

1. **Support** : $\text{Supp}(X \rightarrow Y) = p(X' \cap Y')$
   $\text{Supp}_{inc}(X \rightarrow Y) = 0$, $\text{Supp}_{ind}(X \rightarrow Y) = p(X')p(Y')$, $\text{Supp}_{imp}(X \rightarrow Y) = p(X')$, donc $det(M) = p^2(X')p(\overline{X})p(Y')p(\overline{Y}') \neq 0$. D'après le Théorème 21, la mesure Support est normalisable.

   $$x_f = \frac{1}{p(X')(1-p(Y'))}, \quad y_f = -\frac{p(X')p(Y')}{p(X')(1-p(Y'))}$$
   $$x_d = \frac{1}{p(X')p(Y')}, \quad y_d = -1$$

   soit

   $$\text{Supp}_n(X \rightarrow Y) = \begin{cases} \frac{p(X' \cap Y') - p(X')p(Y')}{p(X')(1-p(Y'))} & \text{si } X \text{ favorise } Y \\ \frac{p(X' \cap Y') - p(X')p(Y')}{p(X')p(Y')} & \text{si } X \text{ défavorise } Y \end{cases}$$

   Finalement, on trouve que $\text{Supp}_n(X \rightarrow Y) = \text{M}_{\text{GK}}(X \rightarrow Y)$.

2. **Confiance** : $\text{Conf}(X \rightarrow Y) = p(Y'|X')$
   $\text{Conf}_{inc}(X \rightarrow Y) = 0$, $\text{Conf}_{ind}(X \rightarrow Y) = p(Y')$ et $\text{Conf}_{imp}(X \rightarrow Y) = 1$, donc $det(M) = 1 - p(Y') \neq 0$.
   D'après le Théorème 21, la mesure Confiance est normalisable.

   $$x_f = \frac{1}{1-p(Y')}, \quad y_f = -\frac{p(Y')}{1-p(Y')}$$
   $$x_d = \frac{1}{p(Y')}, \quad y_d = -1$$

   soit

   $$\text{Conf}_n(X \rightarrow Y) = \begin{cases} \frac{p(X' \cap Y') - p(X')p(Y')}{p(X')(1-p(Y'))} & \text{si } X \text{ favorise } Y \\ \frac{p(X' \cap Y') - p(X')p(Y')}{p(X')p(Y')} & \text{si } X \text{ défavorise } Y \end{cases}$$

   On trouve que $\text{Conf}_n(X \rightarrow Y) = \text{M}_{\text{GK}}(X \rightarrow Y)$.

3. **Lift** : $Lift(X \rightarrow Y) = \frac{p(X' \cap Y')}{p(X')p(Y')}$
   $Lift_{inc} = 0$, $Lift_{ind} = 1$, $Lift_{imp} = \frac{1-p(Y')}{p(Y')}$, donc $det(M) = \frac{1-p(Y')}{p(Y')} \neq 0$.
   La mesure de qualité $Lift$ est donc normalisable.

   $$x_f = \frac{p(Y')}{1-p(Y')}, \quad y_f = -\frac{p(Y')}{1-p(Y')}$$
   $$x_d = 1, \quad y_d = -1$$

soit

$$Lift_n(X{\rightarrow}Y) = \begin{cases} \frac{p(X'\cap Y')-p(X')p(Y')}{p(X')(1-p(Y'))} \text{ si } X \text{ favorise } Y \\ \frac{p(X'\cap Y')-p(X')p(Y')}{p(X')p(Y')} \text{ si } X \text{ défavorise } Y \end{cases}$$

On trouve que $Lift_n(X{\rightarrow}Y) = \mathrm{M_{GK}}(X{\rightarrow}Y)$.

4. **Laplace** : $Lap(X{\rightarrow}Y) = \frac{np(X'\cap Y')+1}{np(X')+2}$
$Lap_{inc} = \frac{1}{np(X')+2}$, $Lap_{ind} = \frac{np(X')p(Y')+1}{np(X')+2}$, $Lap_{imp} = \frac{np(X')+1}{np(X')+2}$, donc
$det(M) = \frac{n^2p^2(X')p(Y')(1-p(Y'))}{(np(X')+2)^2} \neq 0$.
La mesure de qualité Lap est donc normalisable.

$$\begin{aligned} x_f &= \frac{np(X')+2}{np(X')(1-p(Y'))}, \quad y_f = -\frac{np(X')p(Y')+1}{np(X')(1-p(Y')} \\ x_d &= \frac{np(X')+2}{np(X')p(Y')}, \quad y_d = -\frac{np(X')p(Y')+1}{np(X')p(Y')} \end{aligned}$$

soit

$$Lap_n(X{\rightarrow}Y) = \begin{cases} \frac{p(X'\cap Y')-p(X')p(Y')}{p(X')(1-p(Y'))} \text{ si } X \text{ favorise } Y \\ \frac{p(X'\cap Y')-p(X')p(Y')}{p(X')p(Y')} \text{ si } X \text{ défavorise } Y \end{cases}$$

On trouve que $Lap_n(X{\rightarrow}Y) = \mathrm{M_{GK}}(X{\rightarrow}Y)$.

5. ϕ-**coefficient** : $\phi(X{\rightarrow}Y) = \frac{p(X'\cap Y')-p(X')p(Y')}{\sqrt{p(X')p(Y')p(\overline{X}')p(\overline{Y}')}}$
$\phi_{inc} = -\sqrt{\frac{p(X')p(Y')}{p(\overline{X}')p(\overline{Y}')}}$, $\phi_{ind} = 0$, $\phi_{imp} = \sqrt{\frac{p(X')p(Y')}{p(\overline{X}')p(\overline{Y}')}}$, donc
$det(M) = -\frac{p(X')p(Y')}{(p(\overline{X}')p(\overline{Y}')} \neq 0$.
La mesure de qualité ϕ est donc normalisable.

$$\begin{aligned} x_f &= \frac{\sqrt{p(X')p(Y')p(\overline{X}')p(\overline{Y}')}}{p(X')(1-p(Y'))}, \quad y_f = 0 \\ x_d &= \frac{\sqrt{p(X')p(Y')p(\overline{X}')p(\overline{Y}')}}{p(X')p(Y')}, \quad y_d = 0 \end{aligned}$$

soit

$$\phi_n(X{\rightarrow}Y) = \begin{cases} \frac{p(X'\cap Y')-p(X')p(Y')}{p(X')(1-p(Y'))} \text{ si } X \text{ favorise } Y \\ \frac{p(X'\cap Y')-p(X')p(Y')}{p(X')p(Y')} \text{ si } X \text{ défavorise } Y \end{cases}$$

On trouve que $Lap_n(X{\rightarrow}Y) = \mathrm{M_{GK}}(X{\rightarrow}Y)$.

5. **Jaccard** : $Jac(X \rightarrow Y) = \frac{p(X' \cap Y')}{p(X')+p(Y')-p(X' \cap Y')}$
   $Jac_{inc} = 0$, $Jac_{ind} = \frac{p(X')p(Y')}{p(X')p(\overline{Y}')+p(Y')}$, $Jac_{imp} = \frac{p(X')}{p(Y')}$, donc $det(M) \neq 0$.
   La mesure de qualité Jac est donc normalisable.

$$x_f = \frac{p(Y')(p(X')p(\overline{Y}')+p(Y'))}{p(X')p(\overline{Y}')(p(X')+p(Y'))}, \quad y_f = -\frac{p^2(Y')}{p(\overline{Y}')(P(X')+p(Y')}$$
$$x_d = \frac{p(X')p(\overline{Y}')+p(Y')}{p(X')p(Y')}, \quad y_d = -1$$

   soit

$$Jac_n(X \rightarrow Y) = \begin{cases} \frac{p(Y')(p(X')p(\overline{Y}')+p(Y'))}{p(X')p(\overline{Y}')(p(X')+p(Y'))} Jac(X \rightarrow Y) - \frac{p^2(Y')}{p(\overline{Y}')(P(X')+p(Y')} \\ \text{si } X \text{ favorise } Y \\ \frac{p(X')p(\overline{Y}')+p(Y')}{p(X')p(Y')} Jac(X \rightarrow Y) - 1 \text{ si } X \text{ défavorise } Y \end{cases}$$

   On constate que $Jac_n(X \rightarrow Y) \neq \mathrm{M_{GK}}(X \rightarrow Y)$.

6.5 Classification de mesures de qualité des règles

Des études formelles sur les mesures de qualité ont permis de faire des classifications de ces mesures. Lallich et Teytaud [LT04] catégorisent les mesures de qualité en deux classes : mesures statistiques et mesures descriptives. Une mesure de qualité est dite mesure *statistique* si elle varie en fonction de la taille de données. Le Support, le Lift, la Conviction sont des exemples des mesures statistiques. Une mesure de qualité est dite mesure *descriptive* si elle est insensible à la taille de données. La Confiance, la Spécificité, la Fiabilité négative, $\mathrm{M_{GK}}$ sont des exemples des mesures de qualité descriptives. Blanchard et al. [BGBG05] proposent une autre méthode de classifications des mesures de qualité : mesures de déviation d'indépendance et mesures de déviation d'équilibre. Une mesure est dite mesure de déviation d'indépendance, si elle prend une valeur constante en cas d'indépendance entre la prémisse et le conséquent. La Conviction, la Satisfaction, le Lift, la J-mesure, l'indice de Loevinger sont des exemples des mesures de déviation d'indépendance. Une mesure est dite mesure *de déviation d'équilibre* si elle prend une valeur constante en cas d'équilibre, *i.e.*, si le nombre d'exemples et de contre-exemples de la règle sont égaux. La Confiance, la mesure de Sebag, le Moindre-contradiction, la Spécificité etc. sont des mesures de déviation d'équilibre. Au sens de la normalisation présentée dans le présent chapitre, nous proposons une nouvelle méthode classification des mesures de qualité

des règles. Nous montrons que ces mesures peuvent être catégorisées en trois classes :

(i) mesures M_{GK}-normalisables ;

(ii) mesures normalisables à normalisées différentes de M_{GK} ;

(iii) mesures non normalisables.

Numéro	Measure	Expression	Reférence
1	Support	$p(X' \cap Y')$	[AIS93]
2	Confiance	$p(Y'\|X')$	[AIS93]
3	M_{GK}	$\frac{p(Y'\|X')-p(Y')}{1-p(Y')}$ *si* $p(Y'\|X') \geq p(Y')$ $\frac{p(Y'\|X')-p(Y')}{p(Y')}$ *si* $p(Y'\|X') \leq p(Y')$	[Gui00]
4	Rappel	$p(X'\|Y')$	[LFZ99]
5	Lift	$\frac{p(Y'\|X')}{p(Y')}$	[BMS97]
6	Laverage	$p(Y'\|X') - p(X')p(Y')$	cf. [GH06]
7	Confiance centrée	$p(Y'\|X') - p(Y')$	[LT04]
8	Facteur de certitude	$\frac{p(Y'\|X')-p(Y')}{1-p(Y')}$	cf. [GH06]
9	Laplace	$\frac{n.p(X'\cap Y')+1}{np(X')+2}$	[Goo65]
10	ϕ-coefficient	$\frac{p(X'\cap Y')-(X')p(Y')}{\sqrt{p(X')p(Y')p(\overline{X}')p(\overline{Y}')}}$	[Ler81]
11	Piatetsky-Shapiro	$p(X' \cap Y') - p(X')p(Y')$	[PS91]
12	Cosinus	$\frac{p(X'\cap Y')}{\sqrt{p(X')p(Y')}}$	cf. [HGB05a]
13	Accuracy	$P(X' \cap Y') + p(\overline{X}' \cap \overline{Y}')$	cf. [GH06]
14	Moindre Contradiction	$\frac{p(X'\cap Y')-p(\overline{Y}\cap\overline{Y}')}{p(Y')}$	[AK02]
15	Loevinger	$1 - \frac{p(X'\cap\overline{Y}')}{p(X')p(\overline{Y}')}$	[Loe47]
16	Kappa	$2\frac{p(X'\cap Y')-p(X')p(Y')}{p(X')+p(Y')-2p(X')p(Y')}$	[Coh60]
17	Indice d'Implication	$\sqrt{n}\frac{p(X'\cap Y')-p(X')p(Y')}{\sqrt{p(X')p(Y')}}$	[LGR81]
18	Spécificité	$p(\overline{Y}'\|\overline{X}')$	[LFZ99]
19	Fiabilité Négative	$p(\overline{X}'\|\overline{Y}')$	[LFZ99]

Tab. 6.4: *Mesures de qualité* M_{GK}*-normalisables*

6.5.1 Mesures M_{GK}-normalisables

Cette première classe contient la plupart des mesures de qualité proposées dans la littérature. Ces mesures sont présentées dans le Tableau 6.4. Elles

Numéro	Mesure de qualité	Coefficient x_f	coefficient y_f
1	Support	$\frac{1}{p(X')(1-p(Y'))}$	$-\frac{p(Y')}{1-p(Y')}$
2	Confiance	$\frac{1}{1-p(Y')}$	$-\frac{p(Y')}{1-p(Y')}$
3	M_{GK}	1	0
4	Rappel	$\frac{p(Y')}{p(X')(1-p(Y'))}$	$-\frac{p(Y')}{1-p(Y')}$
5	Lift	$\frac{p(Y')}{1-p(Y')}$	$-\frac{p(Y')}{1-p(Y')}$
7	Laverage	$\frac{1}{1-p(Y')}$	$-\frac{p(\overline{X}')p(Y')}{1-p(Y')}$
8	Confiance-centrée	$\frac{1}{1-p(Y')}$	0
9	Facteur de certitude	1	1
10	Laplace	$\frac{np(X')+2}{np(X')(1-p(Y'))}$	$-\frac{np(X')p(Y')+1}{np(X')(1-p(Y'))}$
10	ϕ-coefficient	$\frac{\sqrt{p(X')p(Y')p(\overline{X}')p(\overline{Y}')}}{p(X')(1-p(Y'))}$	0
11	Piatetsky-Shapiro	$\frac{1}{p(X')(1-p(Y'))}$	0
12	Cosinus	$\frac{\sqrt{p(Y')}}{\sqrt{p(X')(1-p(Y'))}}$	$-\frac{p(Y')}{1-p(Y'))}$
13	Accuracy	$\frac{1}{2p(X')(1-p(Y'))}$	$\frac{p(X')p(\overline{Y}')+p(\overline{X}')p(Y')-1}{2p(X')(1-p(Y'))}$
14	Moindre Contradiction	$\frac{p(Y')}{2p(X')(1-p(Y'))}$	$-\frac{2p(X')p(Y')-p(X')}{2p(X')(1-p(Y'))}$
15	Loevinger	1	0
16	Kappa	$\frac{p(X')+p(Y')-2p(X')p(Y')}{2p(X')(1-p(Y'))}$	0
17	Indice d'Implication	$-\frac{1}{\sqrt{np(X')(1-p(Y'))}}$	0
18	Spécificité	$\frac{1-p(X')}{p(X')(1-p(Y'))}$	$-\frac{1-p(X')-p(Y')+p(X')p(Y')}{p(X')(1-p(Y'))}$
19	Fiabilité Négative	$\frac{1}{p(X')}$	$-\frac{1-p(X')-p(Y')+p(X')p(Y')}{p(X')(1-p(Y'))}$

Tab. 6.5: *Coefficients de normalisation des mesures* M_{GK}*-normalisables dans le cas où la prémisse favorise le conséquent.*

Numéro	Mesure de qualité	Coefficient x_d	Coefficient y_d
1	Support	$\frac{1}{p(X')p(Y')}$	-1
2	Confiance	$\frac{1}{p(Y')}$	-1
3	M_{GK}	1	0
4	Rappel	$\frac{1}{p(X')}$	-1
5	Lift	1	-1
6	Laverage	$\frac{1}{p(Y')}$	$-p(\overline{X}')$
7	Confiance-centrée	$\frac{1}{p(Y')}$	0
8	Facteur de certitude	$\frac{1-p(Y')}{p(Y')}$	0
9	Laplace	$\frac{np(X')+2}{np(X')p(Y')}$	$-\frac{np(X')p(Y')+1}{np(X')p(Y')}$
10	ϕ-coefficient	$\frac{\sqrt{p(X')p(Y')p(\overline{X}')p(\overline{Y}')}}{p(X')p(Y')}$	0
11	Piatetsky-Shapiro	$\frac{1}{p(X')p(Y')}$	0
12	Cosinus	$\frac{1}{\sqrt{p(X')p(Y')}}$	-1
13	Accuracy	$\frac{1}{p(X')p(Y')}$	$\frac{1-p(X')p(\overline{Y}')-p(\overline{X}')p(Y')}{2p(X')p(Y')}$
14	Moindre Contradiction	$\frac{1}{2p(X')}$	$-\frac{2p(Y')-1}{2p(Y')}$
15	Loevinger	$\frac{1-p(Y')}{p(Y)}$	0
16	Kappa	$\frac{p(X')+p(Y')-2p(X')p(Y')}{2p(X')p(Y')}$	0
17	Indice d'Implication	$\frac{\sqrt{p(X')(1-p(Y'))}}{\sqrt{n}p(X')p(Y')}$	0
18	Spécificité	$\frac{1-p(X')}{p(X')p(Y')}$	$-\frac{1-p(X')-p(Y')+p(X')p(Y')}{p(X')p(Y')}$
19	Fiabilité Négative	$\frac{1-p(Y')}{p(X')p(Y')}$	$-\frac{1-p(X')-p(Y')+p(X')p(Y')}{p(X')p(Y')}$

Tab. 6.6: *Coefficients de normalisation des mesures M_{GK}-normalisables dans le cas où la prémisse défavorise le conséquent.*

peuvent s'écrire sous forme de fonction affine par morceaux de la mesure de qualité M_{GK}, avec des coefficients variables ou dynamiques. Les coefficients de transformation dans le cas où la prémisse favorise le conséquent sont donnés dans le Tableau 6.5, et ceux dans le cas où la prémisse défavorise le conséquent sont présentés dans le Tableau 6.6. Notons que les règles positives correspondent au premier cas, et que les règles négatives au second cas.

6.5.2 Mesures normalisables à normalisées différentes de M_{GK}

Ces mesures vérifient la condition nécessaire et suffisante pour qu'une mesure qualité soit normalisable, mais leurs normalisées associées ne sont pas M_{GK}. La mesure Jaccard, la mesure de Zhang sont des mesures normalisables dont leurs normalisées sont différentes de M_{GK}. Le Tableau 6.7 présente une liste des mesures normalisables dont leurs normalisées associées est différentes de M_{GK}.

Mesure de qualité	Expression	Référence
Jaccard	$\frac{p(X'\cap Y')}{p(X')+p(Y')-p(X'\cap Y')}$	[Jac08]
Zhang	$\frac{p(X'\cap Y')-p(X')p(Y')}{max\{p(X'\cap Y')p(\overline{Y}');p(Y')p(X'\cap\overline{Y}')\}}$	[Zha00]
Q-Yule	$\frac{p(X'\cap Y')p(\overline{X'\cap Y'}-p(X'\cap\overline{Y}')p(\overline{X}'\cap Y')}{p(X'\cap Y')p(\overline{X'\cap Y'})+p(\overline{X}'\cap Y')p(X'\cap\overline{Y}')}$	cf [GH06]
Y-Yule	$\frac{\sqrt{p(X'\cap Y')p(\overline{X'\cap Y'})}-\sqrt{p(X'\cap\overline{Y}')p(\overline{X}'\cap Y')}}{\sqrt{p(X'\cap Y')p(\overline{X'\cap Y'})}+\sqrt{p(X'\cap\overline{Y}')p(\overline{X}'\cap Y')}}$	cf[GH06].
J-mesure	$p(X'\cap Y')log(\frac{p(X'\cap Y')}{p(X')p(Y')})$ $+p(X'\cap\overline{Y}')log(\frac{p(X'\cap\overline{Y}')}{p(X')p(\overline{Y}')})$	[GS88]

Tab. 6.7: *Mesures normalisables à normalisée différente de* M_{GK}

6.5.3 Mesures non normalisables

Ces mesures sont des mesures qui ne vérifient pas la condition nécessaire et suffisante de la normalisabilité de mesure de qualité. Elles ont une valeur infinie en un point de référence (par exemple, le cas de la mesure Conviction) ou elles prennent une même valeur en deux points de référence (par exemple,

le cas de la mesure de Klosgen). Une liste de mesures non normalisables est présentée dans le Tableau 6.8.

Mesure	Expression	Justification
Multiplicateur de cote	$\frac{p(X'\cap Y').p(\overline{Y'})}{p(X'\cap\overline{Y'})p(Y')}$	$\mu_{imp}(X\rightarrow Y) = +\infty$
Sebag	$\frac{p(Y'/X')}{p(\overline{Y}/X')}$	$\mu_{imp}(X\rightarrow Y) = +\infty$
Conviction	$\frac{p(X').p(\overline{Y'})}{p(X'\cap\overline{Y'})}$	$\mu_{imp}(X\rightarrow Y) = +\infty$
Odd Ratio	$\frac{p(X'\cap Y').p(\overline{X'}\cap\overline{Y'})}{p(\overline{X'}\cap Y')p(X'\cap\overline{Y'})}$	$\mu_{imp}(X\rightarrow Y) = \infty$
Klosgen	$\sqrt{p(X'\cap Y')}(p(Y'/X') - p(Y'))$	$\mu_{ind}(X\rightarrow Y) = \mu_{inc}(X\rightarrow Y)$
Gain Informationnel	$log\frac{p(X'\cap Y')}{p(X')p(Y')}$	$\mu_{inc} = -\infty$
Exemples contre-exemples	$1 - \frac{p(X'\cap\overline{Y'})}{p(X'\cap Y')}$	$\mu_{inc} = -\infty$

Tab. 6.8: *Liste de mesures non normalisables*

6.6 Conclusion

Nous avons étudié les mesures de qualité des règles d'association au sens de la normalisation de ces mesures. Ici normaliser une mesure de qualité des règles d'association signifie qu'on transporte (si possible) les valeurs prises par cette mesure de qualité sur l'intervalle $[-1, 1]$ par le biais d'une transformation affine de telle sorte que la valeur -1 correspond à l'incompatibilité, les valeurs strictement comprises entre -1 et 0 correspondent à la répulsion, la valeur 0 correspond à l'indépendance, les valeurs strictement comprises entre 0 et 1 correspondent à l'attraction et la valeur 1 correspond à l'implication logique entre la prémisse et le conséquent d'une règle d'association. Nous avons caractérisé les mesures normalisables. Au sens de la normalisation proposée dans le présent travail, nous pouvons catégoriser les mesures de qualité des règles proposées dans la littérature en trois classes :

(i) les mesures dont la normalisée associée est M_{GK} que nous appelons les mesures M_{GK}-normalisables ;

(ii) les mesures normalisables dont leurs normalisées associées sont diférentes M_{GK} ;

(iii) les mesures non normalisables.

Notons que la catégorie (i) contient la plupart de mesures de qualité proposées dans la littérature ; il s'ensuit que la plupart des mesures de qualité sont comparables via leur normalisée commune qu'est M_{GK}. Cette normalisation et la mesure M_{GK} permettent ainsi une vue unificatrice de mesure M_{GK}-normalisables. Les résultats obtenus dans le présent travail nous amène à refléchir sur les questions suivantes :

- Comment rendre normalisables les mesures qui ne sont pas normalisables dans notre approche ?
- Comment exploiter ces résultats à l'extraction des règles d'association d'un contexte de la fouille de données ?

7. BASES POUR LES RÈGLES D'ASSOCIATION

7.1 Introduction

Le présent chapitre concerne des caractérisations de bases pour les règles d'association valides au sens des mesures de qualité Confiance et M_{GK}. Le problème de la pertinence et de l'utilité des règles extraites est un problème majeur de la fouille des règles d'association d'un contexte binaire. Ce problème est lié au nombre des règles d'association extraites qui est en général très important et à la présence d'une forte proportion de règles redondantes, *i.e.*, de règles convoyant la même information parmi celles-ci. L'idée de trouver une base (*i.e.*, un ensemble minimal des règles d'association à partir duquel, on peut dériver toutes les règles valides par utilisation d'un ensemble d'axiomes d'inférence) permet de pallier ces problèmes. Il existe différentes caractérisations de bases, pour les règles d'association valides au sens de la mesure de qualité Confiance, proposées dans la littérature [GD86, Lux91, Pas00b, CS02]. Des propriétés intéressantes de la mesure de qualité M_{GK}; telles que M_{GK} est normalisée (cf. Définition 48 au Chapitre 6) et M_{GK} est la normalisée associée à la plupart des mesures de qualité proposées dans la littérature (cf. Chapitre 6), nous conduit aux études de caractérisation de la bases au sens de cette mesure.

Ce chapitre est organisé de la façon suivante. Nous présentons quelques bases pour les règles d'association valides au sens de la mesure de qualité Confiance dans la Section 7.2. Nous définissons une base pour les règles d'association valides au sens de la mesure de qualité M_{GK} dans la Section 7.3. Cette base est constituée de quatre sous-bases à savoir 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.

7.2 Bases pour les règles Confiance-valides

Plusieurs caractérisations de bases, pour les règles d'association valides au sens de la mesure de qualité Confiance, ont été proposées dans la littérature [Pas00b, PBTL99b, STB+01, CS02]. Dans la présente section, nous présentons quelques unes d'entre elles. En géneral, ces différentes méthodes de caractérisation ont un point commun à savoir classification des règles selon deux types :

- Règles d'association *exactes*, *i.e.*, ce sont des règles d'implications totales. Formellement, les règles d'association exactes sont des règles $X \rightarrow Y$ telles que $X' \subseteq Y'$.
- Règles d'association *approximatives*, *i.e.*, ce sont des règles $X \rightarrow Y$ qui présentent des contre-exemples. Formellement, les règles d'association approximatives sont des règles telles que $X' \nsubseteq Y'$.

Ces deux types de règles peuvent être caractérisés à l'aide de la mesure de qualité utilisée (Confiance, $\mathrm{M_{GK}}$). La proposition ci-dessous caractérise les règles d'association exactes (resp. approximatives) en utilisant la mesure de qualité Confiance.

Proposition 36: Soient X et Y deux motifs d'un contexte de la fouille de données $\mathbb{K}$. La règle d'association $X \rightarrow Y$ est une règle exacte (resp. approximative) si, et seulement si $\mathrm{Conf}(X \rightarrow Y) = 1$(resp. $\mathrm{Conf}(X \rightarrow Y) < 1$).

7.2.1 Base de Guigues-Duquenne-Luxenburger

La base de Guigues-Duquenne pour les règles exactes est définie à partir de l'ensemble des motifs critiques et leurs fermetures selon la fermeture de correspondance de Galois φ.

Définition 49: La base de Guigues-Duquenne [GD86] pour les implications totales (ou logiques) est l'ensemble BDG défini par

$$BDG = \{X \rightarrow \varphi(X) \setminus X : X \ \varphi - \text{critique}\}.$$

Rappelons qu'un motif X est dit φ-critique s'il n'est pas fermé et $\varphi(Y) \subset X$ pour tout Y motif φ-critique strictement contenu dans X [CM03]. Les motifs φ-critiques sont connus aussi sous le nom de pseudo-intensions [GW99].

L'adaptation de la base de Guigues-Duquenne dans le cadre des règles d'association nécessite la prise en compte du Support des motifs critiques et des motifs fermés. Ainsi, la base considérée est la restriction de la base de Guigues-Duquenne sur l'ensemble des motifs fréquents. Nous avons l'expression de la base de Guigues-Duquenne pour les règles d'association exactes (Support, Confiance)-valides.

$$BDG = \{X \rightarrow \varphi(X) \setminus X : X \ \varphi - \text{critique}, \ X \text{ fréquent}\}.$$

Exemple 15: La base de Guigues-Duquenne, pour les règles d'association exactes Confiance-valides, extraite du contexte du Tableau 7.1 est $BDG = \{A \rightarrow C, B \rightarrow E, E \rightarrow B\}$ (avec minsupp $= \frac{2}{6}$).

	A	B	C	D	E
1	1	0	1	1	0
2	0	1	1	0	1
3	1	1	1	0	1
4	0	1	0	0	1
5	1	1	1	0	1
6	0	1	1	0	1

Tab. 7.1: *Contexte de la fouille de données*

Luxenburger définit, dans [Lux91], une base pour les règles d'implications partielles à partir de l'ensemble des motifs fermés.

Définition 50: La base pour les implications partielles est l'ensemble LB défini par

$$LB = \{X \rightarrow Y : X, Y \varphi\text{-fermés}, X \prec Y, \text{Conf}(X \rightarrow Y) \geq \text{minconf}\}$$

Étant donné deux motifs φ-fermés X et Y, $X \prec Y$ signifie que $X \subset Y$ et il n'existe pas de motifs Z tel que $X \subset Z \subset Y$; dans ce cas X est dit "couvert" par Y ou Y " couvre " X.

Comme dans le cas des règles exactes, l'adaptation de la base de Luxenburger dans le cadre des règles d'association nécessite la prise en compte du Support des ensembles fermés. Ainsi, la base considérée est la restriction de la base de Luxenburger sur l'ensemble des motifs fréquents. Nous avons l'expression de

la base de Luxenburger pour les règles d'association approximatives valides au sens des mesures de qualité Support et Confiance.

$$LB = \{X \rightarrow Y : X, Y \varphi\text{-fermés fréquents}, X \prec Y, \text{Conf}(X \rightarrow Y) \geq \text{minconf}\}$$

Exemple 16: La base de Luxenburger pour les règles approximatives extraites du contexte du Tableau 7.1 (pour minsupp $= \frac{2}{6}$ et minconf $= \frac{2}{6}$), est $BL = \{C \rightarrow A, C \rightarrow BE,\ AC \rightarrow BE, BCE \rightarrow A\}$

La réunion de la base de Guigues-Duquenne et la base de Luxenburger forme une base, appelée *base de Guigues-Duquenne-Luxenburger*, pour les règles d'association Confiance-valides.

Les axiomes d'inférence de Armstrong [Arm74] définis ci-dessous permettent de dériver toutes les règles (Support, Confiance)-valides

(A1) $X \supseteq Y$ alors $X \rightarrow Y$;

(A2) $X \rightarrow Y$ et $Y \rightarrow Z$ impliquent $X \rightarrow Z$;

(A3) $X \rightarrow Y$ et $Z \rightarrow T$ impliquent $X \cup Z \rightarrow Y \cup T$.

7.2.2 Base générique et Base informative

Dans [Pas00b, Pas00a], l'auteur propose une autre base pour les règles d'association Confiance-valides. Ses principes se reposent sur l'extraction des règles non redondantes minimales, selon la Définition 51 ci-dessous.

Définition 51: Soit $\mathcal{R}$ l'ensemble des règles d'association Confiance-valides extraites d'un contexte $\mathbb{K}$. Une règle d'association $X \rightarrow Y \in \mathcal{R}$ est dite *non redondante minimale* s'il n'existe pas de règle d'association $Z \rightarrow T \in \mathcal{R}$ telle que $\text{Supp}(Z \rightarrow T) = \text{Supp}(X \rightarrow Y)$, $\text{Conf}(Z \rightarrow T) = \text{Conf}(X \rightarrow Y)$ et $Z \subset X, Y \subset T$.

Dans la présente caractérisation de base, les règles d'association sont classées selon deux types : règles exactes et règles approximatives.

La proposition suivante caractérise les règles d'association exactes

Proposition 37: [Pas00a] Soient X et Y deux motifs d'un contexte $\mathbb{K}$. La règle d'association $X \rightarrow Y$ est une règle d'association exacte si $\varphi(X) = \varphi(Y)$ où φ est l'opérateur de fermeture de la correspondance de Galois.

La base générique pour les règles d'association exactes est donnée par la Définition 52 ci-dessous.

Définition 52: La base générique, pour les règles d'association exactes, est définie par :

$$BGen = \{X \rightarrow Y : Y \varphi \text{ -fermé fréquent, } X \in G_Y \text{ et } X \neq Y\}$$

Rappelons que G_Y est l'ensemble des générateurs minimaux d'un motif fermé Y. Un motif X est dit *générateur minimal* de Y si et seulement si $\varphi(X) = Y$ et il n'existe pas $Z \subset X$ tel que $\varphi(Z) = \varphi(X) = Y$.

Exemple 17: La base générique, pour les règles d'association exactes, extraite du contexte de la fouille de données présenté dans le Tableau 7.1 (avec minsupp $= \frac{2}{6}$), est
$BGen = \{A \rightarrow C, B \rightarrow E, E \rightarrow B, AB \rightarrow CE, AE \rightarrow BC, BC \rightarrow E, CE \rightarrow B\}$

L'axiome d'inférence permettant de retrouver toutes les règles exactes est l'axiome E défini par :

(E) $X \rightarrow Y$, pour tous Z, T tels que $\varphi(Z) = \varphi(T) = \varphi(X) = \varphi(Y)$ impliquent $Z \rightarrow T$.

La Proposition 38 ci-dessous caractérise les règles d'association approximatives.

Proposition 38: [Pas00a] Soient X et Y deux motifs d'un contexte de la fouille de données. La règle d'association $X \rightarrow Y$ est une règle approximative si et seulement si $\varphi(X) \subset \varphi(Y)$.

La Définition 53 ci-dessous définit la base informative pour les règles d'association approximatives.

Définition 53: Soient $\mathbb{K}$ un contexte de la fouille de données, minsupp et minconf respectivement les seuils minimaux de Support et de la Confiance. Notons G l'ensemble des générateurs minimaux des motifs fermés extraits du contexte $\mathbb{K}$.
La base informative pour les règles d'association approximatives est définie par :

$$BI = \{X \rightarrow Y : Y \varphi\text{-fermé fréquent, } X \in G, \varphi(X) \prec Y \text{ et } \text{Conf}(X \rightarrow Y) \geq \text{minconf}\}.$$

Exemple 18: La base informative, pour les règles d'association approximatives, extraite du contexte du Tableau 7.1 (avec Supp = $\frac{2}{6}$ et minconf = $\frac{2}{6}$) est $BI = \{A \rightarrow BCE, B \rightarrow CE, C \rightarrow A, C \rightarrow BE, C \rightarrow ABE, E \rightarrow BC, BC \rightarrow AE, CE \rightarrow AB\}$.

Les axiomes d'inférence permettant de dériver toutes les règles d'association approximatives Confiance-valides sont :

(I1) $X \rightarrow Y$ pour tous Z, T tels que $\varphi(Z) = \varphi(X)$ et $\varphi(T) = \varphi(Y)$ impliquent $Z \rightarrow T$;

(I2) $X \rightarrow Y$ et $Y \rightarrow Z$ impliquent $X \rightarrow Z$.

La réunion de deux sous-bases : base générique pour les règles d'association exactes et base informative pour les règles d'association approximatives, forme une base pour les règles d'association (Support,Confiance)-valides.

7.2.3 Couverture informative pour les règles d'association

L. Cristofor et D. Simovici présentent dans [CS02] une autre base, appelée *couverture informative*, pour les règles d'association valides au sens de la mesure de qualité Confiance. Dans cette caractérisation de bases, les auteurs ne font pas la distinction entre les deux types de règles (exactes et approximatives) durant la phase d'extraction. Avant de présenter la base en question, nous présentons quelques définitions et propositions.

Considérons un contexte de la fouille de données $\mathbb{K} = (\mathcal{E}, \mathcal{A}, \mathcal{R})$ et X, Y, Z et T quatre motifs de $\mathbb{K}$. Nous avons la proposition suivante.

Proposition 39: [CS02] Soient $r_1 : X \rightarrow Y$ et $r_2 : Z \rightarrow T$ deux règles d'association. Si $(Z \cup T) \subseteq (X \cup Y)$ et $\text{Supp}(Z) \leq \text{Supp}(X)$ alors $\text{Supp}(r_2) \geq \text{Supp}(r_1)$ et $\text{Conf}(r_2) \geq \text{Conf}(r_1)$.

La Proposition 39 ci-dessus permet définir l'axiome d'inférence (C) défini de la façon suivante.

(C) $X \rightarrow Y$, pour tous Z, T tels que $(Z \cup T) \subseteq (X \cup Y)$ et $\text{Supp}(X) \leq \text{Supp}(Z)$ impliquent $Z \rightarrow T$.

Définition 54: Soient X, Y, Z et T quatre motifs d'un contexte de la fouille de données. Si la règle d'association $Z \rightarrow T$ peut être dérivée de la règle $X \rightarrow Y$ par utilisation de l'axiome d'inférence (C), on dit que $Z \rightarrow T$ est *couvert* par $X \rightarrow Y$ et on écrit $(X \rightarrow Y) \prec (Z \rightarrow T)$.

Définition 55: Soient X, Y, Z et T quatre motifs d'un contexte de la fouille de données. Les règles d'association $X \rightarrow Y$ et $Z \rightarrow T$ sont dits *équipotentes* si $(X \rightarrow Y) \prec (Z \rightarrow T)$ et $(Z \rightarrow T) \prec (X \rightarrow Y)$.

Proposition 40: [CS02] Soient $X \rightarrow Y$ et $Z \rightarrow T$ deux règles d'association. $X \rightarrow Y$ et $Z \rightarrow T$ sont équipotents si et seulement si $(X \cup Y) = (Z \cup T)$ et $\text{Supp}(X) = \text{Supp}(Z)$

La proposition suivante montre que la relation de couverture " $\prec$" est un préordre sur l'ensemble des règles d'association.

Proposition 41: [CS02] La relation "$\prec$" est reflexive, transitive mais elle n'est pas anti-symétrique

Ci-dessous est donné un corollaire de Proposition 41 précédente.

Corollaire 11: [CS02] Soient X, Y, Z, T, U et V six motifs d'un contexte de la fouille de données. Si les deux règles d'association $X \rightarrow Y$ et $Z \rightarrow T$ sont équipotentes et $(X \rightarrow Y) \prec (U \rightarrow V)$ alors $(Z \rightarrow T) \prec (U \rightarrow V)$.

En se basant sur la Proposition 39, L. Cristofor et D. Simovici définissent la notion de couverture de l'ensemble des règles Confiance-valides.

Définition 56: Soit $\mathcal{R}$ l'ensemble de toutes les règles d'association Confiance-valides extraites d'un contexte de la fouille de données. Une couverture de l'ensemble $\mathcal{R}$ est un ensemble minimal $\mathcal{C} \subseteq \mathcal{R}$ tel que toute règle de $\mathcal{R}$-$\mathcal{C}$ peut être dérivée d'une règle d'association de $\mathcal{C}$ par utilisation de l'axiome d'inférence (C). Une règle d'association appartenant à $\mathcal{C}$ s'appelle règle $\mathcal{C}$-*couverture*.

Notons que l'ensemble $\mathcal{R}$ des règles d'association Confiance-valides peut avoir plusieurs couvertures.

Proposition 42: [CS02] Soit $\mathcal{C}$ une couverture de l'ensemble $\mathcal{R}$ telle qu'une règle $X \rightarrow Y$ est équipotente à une règle $Z \rightarrow T$. Alors, l'ensemble $(\mathcal{C}\text{-}\{X \rightarrow Y\}) \cup \{Z \rightarrow T\}$ est une autre couverture de $\mathcal{R}$.

La proposition suivante exprime quelques propriétés importantes d'une couverture de l'ensemble $\mathcal{R}$ des règles d'association Confiance-valides d'un contexte de la fouille de données.

Proposition 43: Soit $\mathcal{C}$ une couverture de l'ensemble $\mathcal{R}$ des règles d'association extraites d'un contexte de la fouille de données. On a :

(1.) Si $X \rightarrow Y$ et $Z \rightarrow T$ deux règles d'association de $\mathcal{C}$, alors $(X \cup Y) \neq (Z \cup T)$;

(2.) Si $X \rightarrow Y \in \mathcal{C}$, alors pour toute règle $Z \rightarrow T \in \mathcal{R}$ telle que $(X \cup Y) = (Z \cup T)$, on a $\text{Supp}(X \rightarrow Y) \leq \text{Supp}(Z \rightarrow T)$ et $\text{Conf}(X \rightarrow Y) \leq \text{Conf}(Z \rightarrow T)$;

(3.) Si $(X \rightarrow Y) \in \mathcal{C}$, alors il n'existe pas de règle $(Z \rightarrow T) \in \mathcal{R}$ telle que $X = Z$ et $Y \subset Z$.

Parmi les couvertures de l'ensemble $\mathcal{R}$ des règles d'association Confiance-valides extraites d'un contexte de la fouille de données, il existe celles qui sont les plus informatives que d'autres, d'où la définition suivante.

Définition 57: Soient $\mathcal{R}$ un ensemble des règles d'association Confiance-valides extraites d'un contexte de la fouille de données et $\mathcal{C}$ une couverture de $\mathcal{R}$. $\mathcal{C}$ est dite *couverture informative* pour l'ensemble $\mathcal{R}$ si pour toute règle d'association $X \rightarrow Y \in \mathcal{C}$ il n'existe pas de règle $Z \rightarrow T$ qui lui est équipotente telle que $X \subset Z$.

Proposition 44: Soient $\mathcal{C}$ une couverture informative et $X \rightarrow Y \in \mathcal{C}$ avec $I = X \cup Y$ l'ensemble d'attributs de cette règle. Alors, la taille de la prémisse X est inférieure ou égale aux tailles des autres règles Confiance-valides de même ensemble d'attributs I.

7.3 Bases pour les règles M_{GK}-valides

Nous avons présenté dans la section précédente différentes bases pour les régles d'association Confiance-valides. Il est à noter que ces bases sont toutes des bases pour les règles d'association positives. Or, certaines applications nécessitent non seulement la découverte des règles positives mais aussi des règles négatives. Par ailleurs, cette mesure Confiance présente des inconvénients, en particulier elle sélectionne des règles incohérentes à la sémantique d'implication " si ... alors ...". Eu égard à ces problématiques, nous proposons d'adopter dans le présent travail l'utilisation de la mesure M_{GK} [Gui00] pour extraire les règles d'association d'un contexte de de la fouille données. La mesure de qualité M_{GK} permet non seulement la facilité

d'extraction des règles négatives, mais aussi l'écartement des règles telles que la prémisse et le conséquent sont indépendants.

Cette section concerne l'extraction des bases pour les règles d'association valides au sens de la mesure de qualité M_{GK} que nous appelons *règles* M_{GK}-*valides*. Contrairement aux bases présentées dans la section précédente, les bases que nous considérons dans la présente section concerne les règles négatives et positives.

Avant de présenter les bases que nous caractérisons dans le présent travail, nous rappelons quelques notions sur les règles d'association.

7.3.1 Rappels sur les notions des règles d'association

Un contexte de la fouille de données binaires est un couple $\mathbb{K} = (\mathcal{E}, \mathcal{A})$ où $\mathcal{E}$ est un ensemble fini d'entités et $\mathcal{A}$ ensemble de variables binaires définies sur $\mathcal{E}$. Les sous-ensembles de $\mathcal{A}$ seront appelés *motifs positifs* ou tout simplement *motifs*. Étant donné un motif positif X de $\mathbb{K}$:

- l'extension du motif X, notée X', sera définie par l'ensembles des entités vérifiant X, *i.e.*, $X' = \{e \in \mathcal{E} : \forall x \in X, x(e) = 1\}$.
- $\overline{X}$ désigne la négation de X, *i.e.*, $\overline{X}(e) = 1$ si et seulement si il existe $x \in X$ tel que $x(e) = 0$. Nous considérons ces négations des motifs que nous qualifierons *motif négatif* pour distinguer ceux dont ils sont la négation. On notera que $\overline{X}' = \mathcal{E} \setminus X'$.

Soit $\mathbb{K}$ un contexte de la fouille de données. Considérons deux motifs U et V positifs et/ou négatifs du contexte de la fouille de donnée $\mathbb{K}$. Une règle d'association est un couple de motifs (U,V), noté $U \rightarrow V$. Le motif U est appelé la prémisse de $U \rightarrow V$ et V son *conséquent*. Dans la suite, considérons deux motifs positifs X et Y. Quatre types de règles peuvent être obtenus à partir de deux motifs positifs X et Y :

(a) une règle dite *positive*, de la forme $X \rightarrow Y$ ou $Y \rightarrow X$;

(b) une règle dite négative à droite, de la forme $X \rightarrow \overline{Y}$ ou $Y \rightarrow \overline{X}$;

(c) une règle dite négative à gauche, de la forme $\overline{X} \rightarrow Y$ ou $\overline{Y} \rightarrow X$;

(d) une règle dite *bilatéralement négative*, de la forme $\overline{X} \rightarrow \overline{Y}$ ou $\overline{Y} \rightarrow \overline{X}$.

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). Nous rappelons que par définition, une mesure de qualité est est une application μ qui associe une valeur réelle à chaque règle d'association.

Soient μ une mesure de qualité des règles d'association et α un réel. Une règle d'association $U \rightarrow V$ sera dite (μ, α)-*valide* (ou tout simplement μ-*valide* ou *valide*) si $\mu(U \rightarrow V) \geq \alpha$.

Dans cette section, nous nous intéressons aux règles d'association valides au sens de la mesure de qualité $\mathrm{M_{GK}}$, *i.e.*, des règles $U \rightarrow V$ telles que $\mathrm{M_{GK}}(U \rightarrow V) \geq \alpha$ pour un réel $\alpha \in [0, 1]$ donné.

Rappelons les propriétés suivantes de la mesure $\mathrm{M_{GK}}$.

Soient X et Y deux motifs positifs. Nous avons les égalités suivantes :

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

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

La propriété (i) signifie que $\mathrm{M_{GK}}$ est implicative et que les règles bilatéralement négatives sont identiques aux règles positives associées (*i.e.*, leurs contraposées respectives). La propriété (ii) signifie que les règles négatives à gauche peuvent être dérivées par celles négatives à droite et vice-versa. Grâce à ces deux propriétés, nous ne considérons dans la suite que deux types des règles d'association, à savoir les règles positives et celles négatives à droite que nous appellerons tout simplement règles négatives.

Soient U et V deux motifs positifs et/ou négatifs d'un contexte de la fouille de données $\mathbb{K}$. Une règle d'association $U \rightarrow V$ sera appelée *exacte* si $U' \subseteq V'$, où U'et V' sont respectivement l'extension de U et V. Dans le cas contraire, elle sera dite *approximative*.

Les deux types de règles que nous considérons (positives et négatives à droite) peuvent être classées selon les quatre catégories suivantes, pour X, Y deux motifs positifs :

- Règles positives (resp. négatives) exactes, *i.e.*, les règles $X \rightarrow Y$ (resp. $X \rightarrow \overline{Y}$) telles que $X' \subseteq Y'$ (resp. $X' \subseteq \overline{Y}'$) ;
- Règles d'association positives (resp. négatives) approximatives, *i.e.*, $X \rightarrow Y$ (resp.$X \rightarrow \overline{Y}$) telles que $X' \nsubseteq Y'$ (resp. $X' \nsubseteq \overline{Y}'$).

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

Proposition 45: Soient X et Y deux motifs positifs d'un contexte $\mathbb{K}$.

- une règle d'association $X \rightarrow Y$ (resp. $X \rightarrow \overline{Y}$) est exacte si, et seulement si $\mathrm{M_{GK}}(X \rightarrow Y) = 1$ (resp. $\mathrm{M_{GK}}(X \rightarrow \overline{Y}) = 1$).
- une règle d'association $X \rightarrow Y$ (resp. $X \rightarrow \overline{Y}$) est approximative si, et seulement si $\mathrm{M_{GK}}(X \rightarrow Y) < 1$ (resp. $\mathrm{M_{GK}}(X \rightarrow \overline{Y}) < 1$).

De la proposition ci-dessus et la définition de règles valides au sens d'une mesure de qualité, nous avons le corollaire suivant caractérisant les règles $(\mathrm{M_{GK}}, \alpha)$-valides pour $\alpha \in [0, 1]$.

Corollaire 12: Soient X et Y des motifs positifs d'un contexte $\mathbb{K}$. Les règles $(\mathrm{M_{GK}}, \alpha)$-valides sont les règles :

(1) positives exactes, *i.e.*, les règles $X \rightarrow Y$ telles que $\mathrm{M_{GK}}(X \rightarrow Y) = 1$;

(2) négatives exactes, *i.e.*, les règles $X \rightarrow \overline{Y}$ telles que $\mathrm{M_{GK}}(X \rightarrow \overline{Y}) = 1$;

(3) positives approximatives, *i.e.*, les règles $X \rightarrow Y$ telles que $\alpha \leq \mathrm{M_{GK}}(X \rightarrow Y) < 1$;

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

Dans la suite, nous considérerons les bases pour ces quatre catégories de règles d'association. La réunion ces bases constitue une base pour les règles d'association $\mathrm{M_{GK}}$-valides.

Notons qu'une règle $X \rightarrow Y$, où X et Y sont des motifs positifs, serait intéressante si X favorise Y et réciproquement. Dans le cas contraire, la règle $X \rightarrow \overline{Y}$ pourrait être intéressante (car, dans ce cas X favorise $\overline{Y}$ et vice-versa).

Étant donné que nous sommes à l'origine de l'extraction de bases pour les règles d'association $\mathrm{M_{GK}}$-valides, nous pouvons fournir les détails sur les démonstrations des propositions et théorèmes.

7.3.2 Base pour les règles positives exactes

Proposition 46: Soient X et Y des motifs tels que $\text{Supp}(X) \neq 0$ et $\text{Supp}(Y) \neq 1$. Les deux conditions suivantes sont équivalentes :

(i) $\text{M}_{\text{GK}}(X \rightarrow Y) = 1$;

(ii) $\text{Conf}(X \rightarrow Y) = 1$.

Démonstration : Soient X et Y des motifs positifs.

$$\begin{array}{rcl} \text{M}_{\text{GK}}(X \rightarrow Y) = 1 & \Leftrightarrow & \frac{p(Y'|X') - p(Y')}{1 - p(Y')} = 1 \\ & \Leftrightarrow & p(Y'|X') - p(Y') = 1 - p(Y') \\ & \Leftrightarrow & p(Y'|X') = 1 \\ & \Leftrightarrow & \text{Conf}(X \rightarrow Y) = 1. \end{array}$$

Ce qui démontre le résultat. □

Donc l'ensemble des règles positives exactes M_{GK}-valides et les règles positives exactes Conf-valides sont identiques. Ainsi, la base de Gugues-Duquenne [GD86] pour les règles positives exactes Conf-valides est une base pour les règles d'association positives exactes M_{GK}-valides.

$$BPE = BDG = \{X \rightarrow \varphi(X) \setminus X : X \ \varphi - \text{critique}\}.$$

relativement aux axiomes d'inférence de Armstrong [Arm74].

L'algorithme de génération de la base de Guigues-Duquenne présenté dans l'algorithme 11 ci-dessous est l'algorithme proposé dans [Pas00a]. Cet algorithme suppose que les ensembles des motifs fréquents et fermés fréquents sont déjà calculés. Il existe plusieurs algorithmes de génération des motifs fermés dans la littérature citons entre autres : CLOSE [PBTL99a], CLOSET [PHM00], CHARM [ZH99], TITANIC [STB+02],PRINCE [HYS05]. Les notations utilisées dans l'algorithme 11 sont présentées dans le Tableau 7.2.

Algorithme 11:
Entrée: *$\mathcal{F}_i$ ensemble des i-motifs fréquents ; $\mathcal{FC}_i$ ensemble de i-motifs fermés fréquents*
Sortie: *BGD la base de Guigues-Duquenne ;*

1: $BGD \leftarrow \{\}$;
2: ***if*** $(\varphi(\emptyset) \neq \emptyset)$ ***then***
3: $\quad BGD \leftarrow BGD \cup \{\emptyset \rightarrow \varphi(\emptyset)\}$

$\mathcal{F}_i$	Les i-motifs fréquents. Chaque élément de $\mathcal{F}_i$ possède deux champs : motifs et Support.
$\mathcal{FC}_i$	Les i-motifs fermés fréquents. Chaque élément de $\mathcal{FC}_i$ possède deux champs : motif et Support.
$\mathcal{CF}_k$	Les i-motifs critiques potentiels.
BGD	Base de Guigues-Duquenne.
Cumul	Union des fermetures des sous-ensembles critiques du i-motif critique fréquent candidat X considéré
k	Taille maximale de motifs fermés fréquents.

Tab. 7.2: *Notations utilisées dans l'algorithme 11*

4: ***end if***
5: ***for all*** $\mathcal{F}_i$, *pour* $i < k$ ***do***
6: $CPF_i \leftarrow \mathcal{F}_i \setminus \mathcal{FC}_i$);
7: ***for all*** $X \in CPF_i$ ***do***
8: $Cumul \leftarrow \emptyset$;
9: ***for all*** $(C \rightarrow \varphi(C) \setminus C) \in BGD$ ***do***
10: ***if*** $C \subseteq X$ ***then***
11: $cumul \leftarrow cumul \cup \varphi(C)$;
12: ***end if***
13: ***end for***
14: ***if*** $cumul \subset X$ ***then***
15: $BGD \leftarrow BGD \cup \{X \rightarrow \varphi(X) \setminus X\}$;
16: ***end if***
17: ***end for***
18: ***end for***
19: *Retourner* BGD ;

L'algorithme commence par initialiser l'ensemble BGD avec l'ensemble vide (ligne 1). Il détermine ensuite si l'ensemble $\emptyset$ est fermé ou non (s'il n'est pas fermé, il est nécessairement critique). Si $\emptyset$ est critique, la règle $\emptyset \rightarrow \varphi(\emptyset)$ est inserée dans BGD (ligne 3). Ensuite, la boucle dans les étapes (lignes 5-18) construit la base de Guigues-Duquenne de façon itérative. Durant une itération i, l'ensemble CFP_i de i-motifs critiques fréquents candidats est initilalisé avec les i-motifs fréquents $X \in \mathcal{F}_i$, qui ne sont pas des i-motifs fermés fréquents (ligne 6). Ensuite, chacun des i-motifs critiques fréquents candidats C est examiné afin de déterminer s'il est critique (ligne 7-16).

Pour cela, l'union des fermetures des sous-ensembles critiques du motif X est calculée dans le motif *cumul* (lignes 8-13). Ces sous-ensembles critiques sont les antécédents C des règles dans la base de Guigues-Duquenne. Si le motif *cumul* est inclus dans le motif X alors X est un motif critique fréquent et la règle $X \rightarrow \varphi(X) \setminus X$ est insérée dans la base BGD (lignes 14-16). La boucle s'arrête lorsque l'ensemble les motifs fréquents a été considéré et l'ensemble BGD retourné par l'algorithme contient toutes les règles de la base de Guigues-Duquenne.

7.3.3 Base pour les règles négatives exactes

Rappelons que les règles négatives que nous considérons sont des règles d'association de la forme $X \rightarrow \overline{Y}$. La Proposition 47 définit leur Support et Confiance.

Proposition 47: Soient X et Y deux motifs. Nous avons les égalités suivantes :

(1.) $\text{Supp}(\overline{X}) = 1 - \text{Supp}(X)$;

(2.) $\text{Supp}(X \rightarrow \overline{Y}) = \text{Supp}(X) - \text{Supp}(X \rightarrow Y)$;

(3.) $\text{Conf}(X \rightarrow \overline{Y}) = 1 - \text{Conf}(X \rightarrow Y)$.

Le résultat de la Proposition 48 ci-dessous permet de caractériser les règles négatives exactes M_{GK}-valides en fonction de Supports des règles positives correspondantes.

Proposition 48: Soient X et Y deux motifs tels que $\text{Supp}(X) \neq 0$ et $\text{Supp}(Y) \neq 0$. Les deux conditions suivantes sont équivalentes :

(i) $\text{M}_{\text{GK}}(X \rightarrow \overline{Y}) = 1$;

(ii) $\text{Supp}(X \rightarrow Y) = 0$.

Démonstration : Soient X et Y deux motifs positifs.

$$\begin{aligned} \mathrm{M_{GK}}(X{\rightarrow}\overline{Y}) = 1 &\Leftrightarrow \tfrac{p(\overline{Y}'|X') - p(\overline{Y}')}{1-p(\overline{Y}')} = 1 \\ &\Leftrightarrow p(\overline{Y}'|X') - p(\overline{Y}') = 1 - p(\overline{Y}') \\ &\Leftrightarrow p(\overline{Y}'|X') = 1 \\ &\Leftrightarrow 1 - p(Y'|X') = 1 \\ &\Leftrightarrow p(Y'|X') = 0 \\ &\Leftrightarrow p(X' \cap Y') = 0, \text{ car } p(Y') \neq 0 \\ &\Leftrightarrow \mathrm{Supp}(X{\rightarrow}Y) = 0 \end{aligned}$$

Ce qui démontre le résultat. □

La Proposition 48 nous conduit à considérer les axiomes d'inférence suivants, pour tous X, Y et Z :

(NE1) $X{\rightarrow}\overline{Y}$ et $\mathrm{Supp}(Y \cup Z) > 0$ impliquent $X{\rightarrow}\overline{Y \cup Z}$;

(NE2) $X{\rightarrow}\overline{Y}, Z \subset X$ et $\mathrm{Supp}(Z \cup Y) = 0$ impliquent $Z{\rightarrow}\overline{Y}$.

Proposition 49: Les axiomes d'inférence $NE1$ et $NE2$ sont corrects pour les règles négatives exactes, *i.e.*, toute règle d'association déduite, par application de (NE1) et (NE2), à partir d'une règle d'association négative exacte est négative exacte.

Démonstration Nous montrons d'abord que $(NE1)$ est correct. Soit $X{\rightarrow}\overline{Y}$ une règle négative exacte, *i.e.*, $\mathrm{M_{GK}}(X{\rightarrow}\overline{Y}) = 1$. Par la Proposition 48, $\mathrm{Supp}(X \cup Y) = 0$. Donc, pour tout motif Z, on a $\mathrm{Supp}(X \cup (Y \cup Z)) = \mathrm{Supp}(X \cup Y \cup Z) = 0$. Par ailleurs, si Z tel que $\mathrm{Supp}(Z \cup Y) > 0$, alors on obtient, encore par la Proposition 48, que $\mathrm{M_{GK}}(X{\rightarrow}\overline{Y \cup Z}) = 1$. Ce qui démontre la correction de $(NE1)$.
Maintenant, montrons que $(NE2)$ est correct. Soit $X{\rightarrow}\overline{Y}$ une règle négative exacte, *i.e.*, $\mathrm{M_{GK}}(X{\rightarrow}\overline{Y}) = 1$. Donc pour tout motif Z tel que $Z \subset X$ on a $\mathrm{Supp}(X) > 0$. Ainsi, si $\mathrm{Supp}(Z \cup Y) = 0$, alors, par la Proposition 48, on a $\mathrm{M_{GK}}(Z{\rightarrow}\overline{Y}) = 1$. Ce qui démontre que $Z{\rightarrow}\overline{Y}$ est une règle négative exacte. □

Le résultat de la Proposition 48 nous conduit à considérer la bordure positive de l'ensemble des motifs de Support non nul $Bd^+(0)$ [MT97] définie par :

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

Remarque 19: Notons que la bordure positive $Bd^+(0)$ est l'ensemble des motifs maximaux de Support non nul. Elle est identique à l'ensemble des motifs fermés maximaux de Support non nul [PBTL99a].

Nous caractérisons maintenant la base que nous proposons pour l'ensemble des règles négatives exactes $\mathrm{M_{GK}}$-valides.

Théorème 22: [FDT06b, DFT06] L'ensemble BNE défini par :

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

est une base pour les règles négatives exactes $\mathrm{M_{GK}}$-valides relativement aux axiomes d'inférence $NE1$ et $NE2$.

Démonstration : Nous commençons par montrer que toute règle négative exacte $\mathrm{M_{GK}}$-valide peut être dérivée de BNE par application de (NE1) et/ou (NE2). Soit $X \rightarrow \overline{Y}$ une règle négative exacte $\mathrm{M_{GK}}$-valide. 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 à 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. □

Exemple 19: La base BNE, pour les règles négatives exactes, extraite du contexte du Tableau 7.1 est $BNE = \{ACD \rightarrow \overline{B}, ACD \rightarrow \overline{E}, ABCE \rightarrow \overline{D}\}$.

La règle $ABCE \rightarrow \overline{D}$ est une règle de la base BNE. Par application des axiomes (NE1) et (NE2), nous pouvons dériver à partir de cette règle les dix règles : $ABCE \rightarrow \overline{AD}$, $ABCE \rightarrow \overline{CD}$, $ABE \rightarrow \overline{ACD}$, $BE \rightarrow \overline{AD}$, $E \rightarrow \overline{AD}$, $B \rightarrow \overline{AD}$, $E \rightarrow \overline{CD}$, $B \rightarrow \overline{AD}$, $E \rightarrow \overline{ACD}$, $B \rightarrow \overline{ACD}$.

Remarque 20:
- Comme la bordure positive $Bd^+(0)$ est identique à l'ensemble des motifs φ-fermés maximaux de Support strictement positif, donc la base BNE, pour les règles négatives exactes, est exprimée en terme de de l'opérateur de fermeture φ.
- Si la règle $X \rightarrow \overline{Y}$ est exacte alors la règle $Y \rightarrow \overline{X}$ l'est aussi, et réciproquement. Toutefois, ces 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 toutes autres règles négatives exactes $\mathrm{M_{GK}}$-valides combinant les motifs X_1, X_2, Y_1, Y_2.

Dans [FDT07], nous proposons un algorithme de génération de la base BNE pour les règles négatives exactes $\mathrm{M_{GK}}$-valides extraite d'un contexte de la fouille de données $\mathbb{K}$. Le pseudo-code de l'algorithme générant la base BNE est présenté dans l'algorithme 12. Le présent algorithme suppose que la bordure positive $Bd^+(0)$ est déjà trouvée. Il existe dans la littérature différents algorithmes permettant de générer la bordure positive ou les fermés maximaux [LK98, Bay98, ZPOL97]. L'algorithme 12 commence par initialiser l'ensemble BNE à l'ensemble vide (ligne 1). Chaque élément X de l'ensemble $Bd^+(0)$ est examiné successivement (lignes 2 à 6). Pour chaque attribut $x \notin X$, la règle $X \rightarrow \overline{x}$ est insérée dans BNE (lignes 3 à 5).

Algorithme 12 (Base Négative Exacte):
Entrée: *$Bd^+(0)$.*
Sortie: *BNE.*

1: $BNE \leftarrow \{\}$
2: ***for all*** $(X \in Bd^+(0))$ ***do***
3: ***for all*** $x \notin X$ ***do***
4: $BNE \leftarrow BNE \cup \{X \rightarrow \overline{x}\}$
5: ***end for***
6: ***end for***
7: *Retourner BNE*

7.3.4 Base pour les règles positives approximatives

Soit $\alpha \in [0, 1]$. Le résultat de la Proposition 50 ci-dessous caractérise les règles positives approximatives $(\mathrm{M_{GK}}, \alpha)$-valides en fonction de leurs Confiances respectives.

Proposition 50: Soient X et Y deux motifs tels que X favorise Y. Les deux conditions ci-dessous sont équivalentes :

(i) $\alpha \leq M_{GK}(X \rightarrow Y) < 1$;

(ii) $\text{Supp}(Y)(1-\alpha)+\alpha \leq \text{Conf}(X \rightarrow Y) < 1$.

Démonstration :

$$\begin{array}{lll} \alpha \leq \text{M}_{\text{GK}}(X \rightarrow Y) < 1 & \Leftrightarrow & \alpha \leq \frac{p(Y'|X')-p(Y')}{1-p(Y')} < 1 \\ & \Leftrightarrow & \alpha(1-p(Y')) \leq p(Y'|X') - p(Y') < 1 - p(Y') \\ & \Leftrightarrow & \alpha(1-p(Y')) + p(Y') \leq p(Y'|X') < 1 - p(Y') + p(Y') \\ & \Leftrightarrow & p(Y')(1-\alpha) + \alpha \leq p(Y'|X') < 1 \\ & \Leftrightarrow & \text{Supp}(Y)(1-\alpha) + \alpha \leq \text{Conf}(X \rightarrow Y) < 1 \end{array}$$

Ce qui démontre le résultat. □

Le résultat de la Proposition 50 nous conduit à considérer 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 $Z \rightarrow T$.

Les deux lemmes suivants seront utiles pour la démonstration de la Proposition 51 et le Théorème 23. Le Lemme 4 montre que le support d'un motif est égal au Support de sa fermeture [PBTL99a].

Lemme 4: Pour tout motif X, on a : $\text{Supp}(\varphi(X)) = \text{Supp}(X)$.

Le Lemme 5 est une caractérisation des opérateurs de fermeture utilisant une propriété dite d'indépendance de chemins [Plo73].

Lemme 5: Une application extensive ϕ sur $\mathcal{P}(\mathcal{A})$, *i.e.* $X \subseteq \phi(X)$, est un opérateur de fermeture sur $\mathcal{P}(\mathcal{A})$ si et seulement si elle vérifie la propriété $\phi(X \cup Y) = \phi(\phi(X) \cup \phi(Y))$, pour tous $X, Y \in \mathcal{P}(\mathcal{A})$.

Proposition 51: L'axiome d'inférence (PA) est correct pour les règles négatives approximatives $(\text{M}_{\text{GK}}, \alpha)$-valides, *i.e.*, toute règle d'association déduite par application de (PA) à partir d'une règle positive approximative $(\text{M}_{\text{GK}}, \alpha)$-valide est positive approximative $(\text{M}_{\text{GK}}, \alpha)$-valide.

Démonstration : Soit $X \rightarrow Y$ une règle positive approximative $(\text{M}_{\text{GK}}, \alpha)$-valide règle d'association, *i.e.*, $\alpha \leq \text{M}_{\text{GK}}(X \rightarrow Y) < 1$. Alors, par la Proposition 50, $\text{Supp}(Y)(1-\alpha)+\alpha \leq conf(X \rightarrow Y) < 1$. Soient Z et T deux motifs tels que $\varphi(X) = \varphi(Z)$ et $\varphi(Y) = \varphi(T)$. Alors, par Lemmes 4 et 5, $\text{Supp}(X \cup Y) = \text{Supp}(\varphi(X \cup$

$Y)) = \text{Supp}(\varphi(\varphi(X) \cup \varphi(Y))) = \text{Supp}(\varphi(\varphi(Z) \cup \varphi(T))) = \text{Supp}(\varphi(Z \cup T)) = \text{Supp}(Z \cup T)$. Par ailleurs, $\text{Conf}(Z \rightarrow T) = \text{Conf}(X \rightarrow Y)$ donc $\text{Supp}(T)(1-\alpha)+\alpha \leq \text{Conf}(Z \rightarrow T) < 1$. Alors, encore par la Proposition 50, $\alpha \leq \text{M}_{\text{GK}}(Z \rightarrow T) < 1$, ce qui démontre que $Z \rightarrow T$ est approximative $(\text{M}_{\text{GK}}, \alpha)$-valide. □

Par ailleurs, nous avons le résultat suivant :

Théorème 23: [FDT06b, DFT06] L'ensemble $BPA(\alpha)$ défini par

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

est une base pour les règles d'association positives approximatives $(\text{M}_{\text{GK}}, \alpha)$-valides, par rapport à l'axiome d'inférence (PA).

Démonstration : Nous commençons par montrer que toute règle positive approximative $(\text{M}_{\text{GK}}, \alpha)$-valide peut être dérivée de $\text{BPA}(\alpha)$ par application de l'axiome (PA). Soit $X \rightarrow Y$ une règle positive approximative $(\text{M}_{\text{GK}}, \alpha)$-valide. Alors, par la Proposition 50, $\text{Supp}(Y)(1-\alpha)+\alpha \leq \text{Conf}(X \rightarrow Y) < 1$. Considérons les deux motifs φ-fermés $Z = \varphi(X)$ et $T = \varphi(Y)$. D'une part, par le Lemme 4, $\text{Conf}(\varphi(X) \rightarrow \varphi(Y)) = \text{Supp}(\varphi(X) \cup \varphi(Y))/\text{Supp}(\varphi(X)) =$
$\text{Supp}(\varphi(\varphi(X) \cup \varphi(Y)))/\text{Supp}(\varphi(X))$ qui, par le Lemme 5, est égale à $\text{Supp}(\varphi(X \cup Y))/\varphi(X)$ et qui, encore par le Lemme 4, est égale à $\text{Supp}(X \cup Y)/sup(X) = \text{Conf}(X \rightarrow Y)$. D'autre part, par le Lemme 4, $\text{Supp}(\varphi(Y) = \text{Supp}(Y))$, donc
$\text{Supp}(\varphi(Y))(1-\alpha)+\alpha) \leq \text{Conf}(\varphi(X) \rightarrow \varphi(Y)) < 1$. Donc, par la Proposition 50, $0 < \text{M}_{\text{GK}}(\varphi(X) \rightarrow \varphi(Y)) < 1$ alors $\varphi(X) \rightarrow \varphi(Y)$ est un élément de $\text{BPA}(\alpha)$. Par ailleurs, l'application de (PA) à $Z \rightarrow T$ donne la règle $X \rightarrow Y$.

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

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

Remarque 21: Dans la pratique, la base considérée est la restriction de la base ainsi définie sur l'ensemble de motifs fréquents. $BPA(\alpha) = \{X \rightarrow Y : \varphi(X) = X, \varphi(Y) = Y, X, Y$ fréquents $\text{Supp}(Y)(1-\alpha)+\alpha \leq \text{Conf}(X \rightarrow Y) < 1\}$.

Exemple 20: La base positive approximative BPA, pour les règles d'association positives approximatives $\mathrm{M_{GK}}$-valides, extraite du contexte du Tableau 7.1 (avec minsupp $= \frac{2}{6}$ $min\mathrm{M_{GK}} = \frac{2}{6}$) est $BPA(\frac{2}{6}) = \{AC \rightarrow ABCE, BE \rightarrow BCE\}$.
$AC \rightarrow ABCE$ est une règle de la base $BPA(\frac{2}{6})$. Par application de l'axiome d'inférence (PA), nous pouvons dériver les neuf règles d'association $A \rightarrow AB$, $A \rightarrow AE$, $A \rightarrow ABC$,$A \rightarrow ACE$, $A \rightarrow ABCE$ $AC \rightarrow AB$, $AC \rightarrow AE$, $AC \rightarrow ACE$, $AC \rightarrow ABCE$.

7.3.5 Base pour les règles négatives approximatives

Notons qu'une règle valide au sens de la mesure de qualité $\mathrm{M_{GK}}$ est nécessairement une règle où la prémisse favorise le conséquent. 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). Les règles négatives approximatives $(\mathrm{M_{GK}}, \alpha)$-valides sont les règles $X \rightarrow \overline{Y}$ telles que $\alpha \leq \mathrm{M_{GK}}(X \rightarrow \overline{Y}) < 1$. La proposition, ci-dessous, caractérise ces règles en fonction de la Confiance des règles positives correspondantes.

Proposition 52: 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 $Z \rightarrow \overline{T}$.

La proposition suivante montre la correction de l'axiome (NA). Elle se démontre de façon analogue à la Proposition 51.

Proposition 53: . L'axiome d'inférence (NA) est correct pour les règles négatives approximatives $(\mathrm{M_{GK}}, \alpha)$-valides, *i.e.*, toute règle d'association déduite par application de (PA) à partir d'une règle négative approximative $(\mathrm{M_{GK}}, \alpha)$-valide est négative approximative $(\mathrm{M_{GK}}, \alpha)$-valide.

Le théorème 24 ci-dessous caractérise la base que nous proposons pour les règles négatives approximatives $(\mathrm{M_{GK}}, \alpha)$-valides. Il se démontre de façon analogue au Théorème 23.

Théorème 24: [FDT06b, DFT06] L'ensemble $BNA(\alpha)$ défini par

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

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

Exemple 21: Considérons encore une fois le contexte présenté dans la Tableau 7.1. Aucune règle négative approximative n'est valide pour un seuil minimum de Support égal à $\frac{2}{6}$ et un seuil minimum de $\mathrm{M_{GK}}$ égal à $\frac{2}{6}$. Pour un seuil minimum de $\mathrm{M_{GK}}$ égal à $\frac{1}{5}$, on a $BNA(\frac{1}{5}) = \{BE \rightarrow \overline{AC}, AC \rightarrow \overline{BE}\}$.

La règle $BE \rightarrow \overline{AC}$ est une règle de la base $BNA(\frac{1}{5})$. Par application de l'axiome d'inférence (NA), nous pouvons dériver les cinq règles d'association : $B \rightarrow \overline{A}, B \rightarrow \overline{AC}, E \rightarrow \overline{A}, E \rightarrow \overline{AC}, BE \rightarrow \overline{A}$.

Un algorithme de génération de bases BNA et BPA respectivement pour les règles négatives et positives approximatives que nous proposons est présenté dans l'algorithme 14.

Remarquons que si X et Y sont deux motifs tels que X et Y sont comparables, (*i.e.*, ou bien $X \subseteq Y$ ou bien $Y \subseteq X$) alors X favorise Y et réciproquement. Cela permet, pour un motif fermé X, de restreindre l'espace de recherche des motifs négatifs conséquents potentiels de X aux fermés incomparables avec X. Avant de présenter le pseudo-code de l'algorithme de construction de bases positive et négative pour les règles approximatives, nous donnons ci-dessous l'algorithme de construction des fermés incomparables à un motif fermé donné. Le pseudo-code de l'algorithme générant les fermés incomparables à un motif fermé X, de taille inférieure au égale à la taille de X, est présenté dans l'algorithme 13. Les notations utilisées dans l'algorithme 13 sont présentées dans le Tableau 7.3.

$NCom_{\leq i}(X)$	Les motifs non comparables à X de taille $\leq \|X\|$
$SEns(\mathcal{FC}_j, X)$	Les fermés à la fois sous-ensembles de X et de $\mathcal{FC}_j$
$NCom_j$	Les fermés non comparables à X de taille j

Tab. 7.3: *Notations utilisées dans l'algorithme 13*

Algorithme 13 ($NCom_{\leq i}(X)$):
Entrée: *$\mathcal{FC}_i$: ensemble des i-motifs fermés.*
Sortie: *Ensemble des fermés non comparables à X de taille $i \leq |X|$.*

1: $NCom_{\leq i}(X) \leftarrow \{\}$
2: ***for all*** $(\mathcal{FC}_j, j \leq |X|)$ ***do***
3: $\quad NCom_j \leftarrow FC_j \setminus SEns(\mathcal{FC}_j, X)$;
4: $\quad NCom_{\leq i}(X) \leftarrow NCom_{\leq i}(X) \cup NCom_j$
5: ***end for***
6: $NCom_{\leq i}(X)$

L'initialisation de $NCom_{\leq i}(X)$ à l'ensemble vide est faite en ligne 1. L'étape dans la ligne 3 calcule les ensembles $NCom_j$ de φ-fermés non comparables avec X, de taille j. $NCom_j$ est obtenu à partir de FC_j en supprimant ses éléments qui sont sous-ensembles de X. L'ensemble des φ-fermés incomparables à X, de taille inférieure ou égale à la taille de X, est obtenu en ligne 4. Les notations utilisées dans l'algorithme 14 sont présentées dans le Tableau 7.4.

$\mathcal{FC}_i$	Les i-motifs fermés fréquents
k	Taille maximale des motifs fermés fréquents
$BNA(resp.BPA)$	Base négative (resp. positive) approximatives
$SEns(\mathcal{FC}_j, Y)$	Les sous-motifs fermés qui sont sous-motifs de Y

Tab. 7.4: *Notations utilisées dans l'algorithme 14*

Algorithme 14 (Bases Négative-Positive Approximatives):
Entrée: *$\mathcal{FC}_i$: ensemble des fermés fréquents et leurs supports ; α : seuil M_{GK}.*
Sortie: *BNA,BPA : respectivement bases pour les règles négatives et positives approximatives valides.*

1: $BNA \leftarrow \{\}; BNA \leftarrow \{\}$;
2: ***for all*** $(\mathcal{FC}_i, i \leq k)$ ***do***
3: ***for all*** $Y \in \mathcal{FC}_i$ ***do***
4: ***for all*** $X \in SEns(\mathcal{FC}_j, Y)$ ***do***
5: Conf $\leftarrow \frac{\text{Supp}(X \cup Y)}{\text{Supp}(X)}$;
6: ***if*** $\text{Supp}(X \rightarrow Y)(1-\alpha) + \alpha \leq \text{Conf} < 1$ ***then***
7: $BPA \leftarrow BPA \cup \{X \rightarrow Y\}$;
8: ***end if***
9: ***end for***
10: ***for all*** $X \in NCom_{\leq i}(Y)$ ***do***
11: Conf $\leftarrow \frac{\text{Supp}(X \cup Y)}{\text{Supp}(X)}$
12: ***if*** $0 < \text{Conf} \leq \text{Supp}(Y)(1-\alpha)$ ***then***
13: $BNA \leftarrow BNA \cup \{X \rightarrow \overline{Y}, Y \rightarrow \overline{X}\}$;
14: ***else***
15: ***if*** $\text{Supp}(Y)(1-\alpha) + \alpha \leq \text{Conf} < 1$ ***then***
16: $BPA \leftarrow BPA \cup \{X \rightarrow Y\}$;
17: ***end if***
18: ***end if***
19: ***end for***

20: ***end for***
21: ***end for***
22: *Retourner* BNA, BPA;

L'algorithme 14 commence par initialiser les bases BNA et BPA à l'ensemble vide (ligne 1). Ensuite, les ensembles $\mathcal{FC}_i$ sont parcourus successivement dans l'ordre décroissant (ou croissant) de i (lignes 2à 21). Durant une itération i, à partir de tout motif fermé Y il y a deux étapes :

(a) l'algorithme construit une partie de la base BPA à partir des fermés qui sont sous motifs de Y vérifiant la condition de validité d'une règle positive approximative (lignes 3 à 9) ;

(b) l'algorithme cherche les motifs fermés X, de taille inférieure à la taille de Y, qui lui sont incomparables et insère dans la base BNA les règles $X \rightarrow \overline{Y}$ vérifiant la condition de validité d'une règle d'association négative approximative. Comme $\mathrm{M_{GK}}(X \rightarrow \overline{Y}) = \mathrm{M_{GK}}(Y \rightarrow \overline{X})$, la règle $Y \rightarrow \overline{X}$ est aussi insérée dans la base BNA. Si la règle $X \rightarrow \overline{Y}$ n'est pas valide, l'algorithme vérifie si la règle positive $X \rightarrow Y$ est valide et l'insère par la suite dans la base BPA dans le cas où elle est valide (lignes 14 à 18).

Enfin, les deux bases BNA et BPA sont construites.

7.3.6 Exemple d'illustration

En guise d'illustration de la méthode ainsi proposée, nous proposons ci-dessous une application sur le jeu de données bancaires que nous empruntons à la thèse de S. Guillaume [Gui00]. En effet, il s'avère opportun ici de faire une comparaison des nombres des règles dans nos bases positives et celles dans la Base de Guigues-Duquenne-Luxenburger à partir de ce petit jeu de données. Le Tableau 7.5. présente la base de données bancaires que nous utiliserons pour illustrer notre méthode. Elle est constituée de 10 individus et quatre variables qu'on a discrétisés en 9 variables binaires selon le Tableau 7.6.

Légendes sur les attributs dans le Tableau 7.6

x_1 : âge $\in]20; 29]$, x_2 : âge $\in]29; 39]$, x_3 : marié, x_4 : profession artiste, x_5 : profession guide, x_6 : profession enseignant, x_7 : mauvais, x_8 : moyen, x_9 : bon.

Le Tableau 7.7 présente les nombres de règles dans les bases ainsi que ceux de toutes les règles $\mathrm{M_{GK}}$-valides, selon les quatre types de règles que nous considérons dans ce travail, extraites à partir du contexte bancaire transformé du Tableau 7.6. Le Tableau 7.7 nous témoigne que les nombres des règles dans les bases sont

Variables Entités	Age	Marié	Profession	Catégorie
e_1	24	oui	artiste	mauvais
e_2	23	non	guide	moyen
e_3	32	oui	enseignant	moyen
e_4	35	oui	artiste	bon
e_5	39	oui	enseignant	bon
e_6	31	oui	artiste	bon
e_7	29	oui	enseignant	bon
e_8	30	oui	enseignant	moyen
e_9	38	oui	enseignant	bon
e_{10}	36	oui	artiste	mauvais

Tab. 7.5: *Données bancaires brutes*

Attributs Entités	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
e_1	1	0	O	1	0	0	1	0	0
e_2	1	0	1	0	1	0	0	0	1
e_3	0	1	1	0	0	1	0	1	0
e_4	0	1	1	1	0	0	0	0	1
e_5	0	1	1	0	0	1	0	1	0
e_6	0	1	1	1	0	0	0	0	1
e_7	1	0	1	0	0	1	0	0	1
e_8	0	1	1	0	0	1	0	1	0
e_9	0	1	1	0	0	1	0	0	1
e_{10}	0	1	1	1	0	0	1	0	0

Tab. 7.6: *Données bancaires transformées*

minsupp	minM$_{GK}$	BNE	RNE	BNA	RNA	BPE	RPE	BPA	RPA
$\frac{2}{10}$	50%	10	180	2	8	7	13	3	24
	40%	10	180	4	12	7	13	3	24
$\frac{3}{10}$	50%	5	59	2	8	5	11	2	20
	50%	5	59	2	8	5	11	2	20

Tab. 7.7: *Nombre de règles dans les bases ainsi que ceux de toutes les règles* M$_{GK}$*-valides*

minsupp	minM$_{GK}$ = minconf	BPA	BL
$\frac{2}{10}$	60%	0	5
	50%	3	10
	40%	3	14
$\frac{3}{10}$	60%	0	5
	50%	2	8
	40%	2	8

Tab. 7.8: *Nombres des règles dans la base BPA et la base de Luxenburger*

significativement très inférieurs aux nombres de toutes les règles M$_{GK}$-valides.

Nous profitons dans le présent travail de faire une comparaison des nombres des règles dans la base BPA pour les règles positives approximatives M$_{GK}$-valides avec celui dans la base de Luxenburger [Lux91] pour les règles positives approximatives Confiance-valides avec les mêmes seuils (*i.e.*, minconf = minM$_{GK}$). Théoriquement M$_{GK}$ est plus sélective que la Confiance. Rappelons que les bases positives exactes M$_{GK}$-valides et Confiances-valides sont identiques. À notre connaissance, les bases pour les règles négatives que nous proposons sont les premières dans la littérature. Le Tableau 7.8 présente les nombres de règles dans la base BPA pour les règles positives approximatives M$_{GK}$-valides et ceux des règles dans la base de Luxenburger pour les règles positives approximatives Confiance-valides extraites à partir du contexte de données bancaires présentées dans le Tableau 7.6.

Ce résultat nous montre bien que le nombre des règles dans la base BPA demeure toujours inférieur au nombre des règles dans la base de Luxenburger. Ceci vient du fait que M$_{GK}$ est plus sélective que la mesure Confiance. Prenons, par exemple, la règles $x_3 \rightarrow x_3x_9 \in BL$ car (Conf($x_3 \rightarrow x_3x_9$) = $\frac{5}{9}$) pour un seuil (pour minsupp = $\frac{3}{10}$ et minconf = 50%) alors que M$_{GK}$($x_3 \rightarrow x_3x_9$) = $\frac{1}{9}$ qui est

une valeur petite, donc la prémisse et le conséquent de cette règle sont voisines de la situation d'indépendance. L'information apportée par cette règle est ainsi négligeable.

7.4 Conclusion

La mesure de qualité Confiance est l'une des mesures les plus utilisées, dans la littérature, pour extraire les règles d'association intéressantes dans un contexte de la fouille de données binaires. Plus souvent, le nombre des règles extraites est très important (nombre exponentiel dans la taille de l'ensemble d'attributs du contexte). Ce qui conduit l'utilisateur à fouiller encore parmi les règles générées. L'extraction de bases pour les règles d'association valides au sens d'une mesure de qualité permet de remédier à ce problème de surabondance sans perte de l'information. Nous avons présenté quelques caractérisations de bases pour les règles d'association valides au sens de la mesure de qualité Confiance. Toutefois, cette mesure de qualité a suscité plusieurs critiques de la part des chercheurs travaillant dans le domaine de la fouille des règles d'association. En effet, elle ne tient pas compte de la situation de référence à l'indépendance et à tort, elle tolère ou produit ainsi la sélection des règles dont la prémisse et le conséquent sont indépendants (les règles de ce type ne portent aucune information nouvelle à l'utilisateur). Et il peut même arriver que des règles qui se trouvent dans la zone de répulsion entre la prémisse et le conséquent (*i.e.*, la prémisse et le conséquent se défavorisent) figurent malhereusement parmi les règles valides. Afin d'éviter une telle incohérence, nous avons proposé dans le présent chapitre d'adopter l'utilisation de la mesure de qualité M_{GK}. Plus particulièrement, nous avons caractérisé des bases au sens de cette mesure de qualité. Les propriétés intéressantes de la mesure de qualité M_{GK} nous permettent de caractériser non seulement des bases pour les règles positives, mais aussi celles pour les règles négatives. Ces bases sont exprimées en termes de l'opérateur de fermeture de connexion de Galois. Nous avons proposé par la suite des algorithmes de construction de ces différentes bases.

Les perspectives des travaux ultérieurs concernent l'optimisation des algorithmes proposés et l'étude des diverses techniques d'implémentation en vue d'appliquer dans des différentes données venant des différents domaines.

8. CONCLUSION ET PERSPECTIVES

Les travaux présentés dans ce mémoire s'inscrivent dans le cadre de l'extraction de connaissances à partir d'un système de bases de données binaires et plus particulièrement dans le domaine de mesures de qualité des règles d'association et caractérisations de bases de ces dernières.

Les méthodes d'extractions automatiques de règles d'association engendrent généralement une quantité prohibitive de règles dont la plupart sont redondantes et faiblement informatives. Cette quantité nuit gravement à l'interprétation des résultats et rend difficile l'analyse et l'identification de celles qui sont réellement intéresantes. Il s'avère donc indispensable d'aider l'utilisateur dans sa recherche des règles intéressantes par utilisation de mesures de qualité pertinentes pour valider l'intérêt des règles.

Des études sur les mesures de qualité des règles ont été effectuées. Nous avons mis en évidence un nouvel éclairage sur l'ensemble de ces mesures. En effet, la normalisation des mesures de qualité des règles permet une classification des différentes mesures de qualité selon trois catégories à savoir les mesures M_{GK}-normalisables, les mesures normalisables dont leurs normalisées associées sont différentes de M_{GK} et les mesures non normalisables. La question qui suit mérite une investigation approfondie :" comment rendre normalisables les mesures qui ne sont pas normalisables dans notre approche ?"

Nous avons caractérisé une base pour les règles d'association valides au sens de la mesure de qualité M_{GK}. Il semble que l'utilisation de cette mesure de qualité permet de pallier les problèmes de méthodes de l'extraction des règles d'association utilisant la mesure de qualité Confiance. Notons par ailleurs que des études des propriétés mathématiques de la mesure M_{GK} montrent que la mesure M_{GK} est plus sélective que la mesure de qualité Confiance. Pour écarter les règles telles que la prémisse et le conséquent sont indépendants, il suffit de fixer le seuil minimum de M_{GK} dans l'intervalle $]0, 1]$. Nous avons proposé par la suite des algorithmes d'extraction de base que nous avons caractérisée. L'implémentation et l'optimisation de ces algorithmes sont les suites naturelles de ces travaux. Enfin, le problème suivant s'avère aussi ouvert : " à partir des caractérisations présentées dans le présent travail, comment caractériser les bases au sens des mesures de qualité M_{GK}-normalisables ?"

À part les questions posées ci-dessus, le présent travail soulève quelques questions méritent davantage d'attention.

- Comment générer efficacement l'ensemble de toutes les règles M_{GK}-valides afin de calculer le taux de réduction apporté par la génération de base ?
- Comment comparer expérimentalement la base proposée dans le présent travail avec la base non redondante de Zaki et al. [Zak00a] ? etc.
- Comment caractériser une base au sens d'une mesure de qualité M_{GK}-normalisable ?
- Dans le présent travail, nous caractérisons une base pour les règles M_{GK}-valides en utilisant les motifs fermés. La question suivante s'avère intéressante. Est-il possible de caractériser une autre base pour les règles M_{GK}-valides en utilisant d'autres représentations condensées des motifs fréquents tels que motifs essentiels, motifs non-dérivables, motifs libres ?

BIBLIOGRAPHIE

[AIS93] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, *Proc. of the ACM SIGMOD International Conference on Management of Data*, volume 22, pages 207–216, Washington,U.S.A., 1993.

[AK02] J. Azé and Y. Kodratoff. Evaluation de résistance au bruit de quelques mesures d'extraction de règles d'association. *Extraction de Connaissances et Apprentissage.*, 14:143–154, 2002.

[Arm74] W. W. Armstrong. Dependency structures of data base relationships. *Information Processing*, 74:580–583, 1974.

[AS94] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. of the 20th VLDB Conference*, pages 487–499, San Diego,Chile, 1994.

[Azé03a] J. Azé. *Extraction de connaissances à partir de données numérique et textuelles*. PhD thesis, Université Paris-Sud XI, Paris,France, 2003.

[Azé03b] J. Azé. Une nouvelle mesure de qualité de l'extraction de petite connaissances. *RSTI série RIA-ECA.*, 17:171–182, 2003.

[Bay98] R. J. Bayardo. Efficiently mining long patterns from databases. In *Proc. of the ACM SIGMOD Conference*, pages 85–93, Washington,U.S.A., June 1998.

[BB00] J. F. Boulicaut and A. Bykowski. Frequent closures as a concise representation for binary data mining. In *Proc. of the Fourth Pacific-Asie Conference on Knowledge Discovery and Data Mining (PAKDD)*, volume 1805 of *LNAI*, pages 62–73, Kyoto,Japan, 2000. Springer-Verlag.

[BBR00] J. F. Boulicaut, A. Bykowski, and C. Rigotti. Approximation of frequency queries by means of free-sets. In *Proc. of the Fourth of*

European Conference Principles and Practice on Knowledge Discovery in Databases (PKDD'00), volume 1910 of *LNAI*, pages 75–85, Lyon,France, 2000. Springer-Verlag.

[BBR03] J. F. Boulicaut, A. Bykowski, and C Rogotti. Free-sets : a condensed representation of boolean data for the approximation of frequency queries. *Data Mining and Knowledge Discovery*, 7(1):5–22, 2003.

[BC07] L. Brisson and M. Collard. Intérêt des systèmes d'information dirigés par ontologies pour la fouille des données, Mars 2007. Rapport de recherche ISRN I3S/RR-2007-08-FR, Centre de Rennes IRISA.

[BGBG05] J. Blanchard, F. Guillet, H. Briand, and R. Gras. Assessing rule with a probabilistic measure of deviation from equilbrium. In *Proc. of 11th International Symposium on Applied Stochastic Models and Data Analysis ASMDA*, pages 191–200, Brest,France, 2005. ENST.

[Bir67] G. Birkhoff. *Lattice theory*. 3rd edition, Coll. Publ., XXV. American Mathematical Society, Providence, RI, 1967.

[BM70] M. Barbut and B. Monjardet. *Ordre et classification*. Hachette, Paris, 1970.

[BMS97] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *Proc. of the ACM SIGMOD Conference*, pages 265–276, Tucson, Arizona, 1997.

[BMUT97] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proc. of the ACM SIGMOD Conference*, pages 255–264, 1997.

[Boo47] G. Boole. *The Mathematical Analysis of Logic*. Cambridge, 1847.

[Bri04] L. Brisson. Mesure d'intérêt subjectif et représentation des connaissances, 2004. Technical Report ISRN I3S/ RR-2004-35FR, Université de Nice.

[BTP+00] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal. Mining frequent patterns with counting inference. *SIGKDD Explorations*, 2:66–75, 2000.

[Byk02] A. Bykowski. *Condensed Representations of Frequent Sets : Application to Descriptive Pattern Discovery*. PhD thesis, Institut National des Sciences Appliquées de Lyon, France, 2002.

[Cal03] T. Calders. Deducing bounds on the support of itemsets. *Database Technologies for Data Mining*, 2682 of LNCS:214–233, 2003.

[Cas99] N. Caspard. A characterization theorem for the canonical basis of a closure operator. *Order*, 16:227–230, 1999.

[CCL05] A. Casali, R. Cicchetti, and L. Lakhal. A perfect cover of frequent patterns. In Lecture Notes in Computer Science, editor, *Proc. of the Data Warehousing and Knowledge Discovery (DaWaK) Conference*, volume 3589, pages 428–437, Copenhagen,Danemark, 2005.

[CG02] T. Calders and B. Goethals. Mining all non-derivable frequent itemsets. In Tapio Elomaa, Heikki Mannila, and Hannu Toivonen, editors, *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*, volume 2431 of *Lecture Notes in Computer Science*, pages 74–85. Springer-Verlag, 2002.

[CG06] T. Calders and B. Goethals. Non-derivable itemset mining. *Data Mining and Knowledge Discovery(DMKD)*, pages 1–35, 2006.

[CM03] N. Caspard and B. Monjardet. The lattices of closure systems, closure operators, and implicational systems on a finite set: a survey. *Discrete Applied Mathematics*, 127:241–269, 2003.

[Coh60] J. Cohen. A coefficient of agreement for nominal scale. *Educational and Psychological Measurement*, 20:37–46, 1960.

[CS96] P. Cheesemean and J. Stutz. Bayesian classification (autoclass): The theory and results. *AAAI Press*, pages 153–180, 1996.

[CS02] L. Cristofor and D. Simovici. Generating an informative cover for association rules, 2002. http://citeseer.nj.nec.com/545875.html.

[Day92] A. Day. The lattice theory of functional dependencies and normal decompositions. *Internat. J. Algebra Comput.*, 2:409–431, 1992.

[Ded00] R. Dedekind. über zelengungen von zahlen durch ihre grösten gemeinsamen teiler. *Gesammelte Werke*, 2:103–148, 1900.

[DFT06] J. Diatta, D. R. Feno, and A. Totohasina. Galois lattices and based for m_{GK}-valid association rules. In *Fourth International Conference on Concept Lattices and Their Applications*, pages 127–138, Tunis, Tunisie, 2006.

[Dia05] J. Diatta. Caractérisation des ensembles critiques d'une famille de moore finie. In *Douzièmes journées de la Société Francophone de Classification*, pages 126–129, Montreal,Canada, 2005.

[DL04] F. Domenach and B. Leclerc. Closure systems, implicational systems, overhanging relations and the case of hierarchical classification. *Mathematical Social Sciences*, 47:349–366, 2004.

[Dom02] F. Domenach. *Structures latticielles, correspondances de Galois contraintes et classification symbolique*. PhD thesis, Université Paris 1 Panthéon-Sorbone, France, 2002.

[DRT07] J. Diatta, H. Ralambondrainy, and A. Totohasina. Towards a unifying probabilistic implicative normalized quality measure for association rules. *Quality Measures in Data Mining*, pages 237–250, 2007.

[EP96] J. Elder and D. Pergibon. A statistical perspective on knowledge discovery in databases. *AAAI Press*, pages 83–115, 1996.

[Eve55] C. J. Everett. Closure operators and Galois Theory in Lattices. *Trans. Amer. Math. Soc.*, 55:514–525, 1955.

[FDT06a] D. R. Feno, J. Diatta, and A. Totohasina. 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.)*, pages 25–30, Lille, France, 2006.

[FDT06b] D. R. Feno, J. Diatta, and A. Totohasina. Une base pour les règles d'association d'un contexte binaire valides au sens de la mesure de qualité m_{GK}. In *Proc. of the 13ème Rencontre de la Société Francophone de Classification*, pages 105–109, Metz, France, 2006.

[FDT07] D. R. Feno, J. Diatta, and A. Totohasina. Génération de bases pour les règles d'association m_{GK}-valides. In *Proc. of the 14ème Rencontre de la Société Francophone de Classification*, pages 101–104, Paris, France, 2007.

[FPSS96] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Knowledge discovery and data mining : towards a unifying framework. In *Proceedings of the second International Conference on Knowledge Discovery and Data Mining*, pages 82–88, Portland, OR, 1996.

[Fre99] A Freitas. On rule interestingness measures. *Knowledge-Based System*, 12:309–315, 1999.

[GD86] J. L. Guigues and V. Duquenne. Famille non redondante d'implications informatives résultant d'un tableau de donnes binaires. *Mathmatiques et Sciences humaines*, 95:5–18, 1986.

[GH06] L. Geng and H. J. Hamilton. Interestingness measures for data mining: A Survey. *ACM Computing Surveys*, 38:1–31, 2006.

[GKCG01] R. Gras, P. Kuntz, R. Couturier, and F. Guillet. Une version entropique de l'intensité d'implication pour les corpus volumineux. *Extraction des Connaissances et Apprentissage*, 1:69–80, 2001.

[Goo65] I.J. Good. The estimation of probabilities: An essay on modern bayesian methods. *The MIT press*, MA, 1965.

[GS88] R. Goodman and P. Smyth. Information theoretic rule induction. In *Proc. of the ECAI-88*, pages 357–362, Munich,Germany, 1988.

[GS97] J. Galambos and I. Simonelli. Bonferroni-type inequalities with applications. *The American Statistical Association*, 440:1649–1650, 1997.

[Gui00] S. Guillaume. *Traitement des données volumineuses. Mesures et algorithmes d'extraction des règles d'association et règles ordinales.* PhD thesis, Universté de Nantes, France, 2000.

[GW99] B. Ganter and R. Wille. *Formal concept analysis, Mathematical foundations.* Springer Verlag, Berlin, 1999.

[GZ01] K. Gouda and M. J. Zaki. Efficiently mining maximal frequent itemsets. In *Proc. of the 1st IEEE International Conference on Data Mining (ICDM)*, pages 163–170, California,U.S.A., 2001.

[HGB05a] X. Huynh, F. Guillet, and H. Briand. Une plateforme exploratoire pour la qualité des règles d'association : Apport pour l'analyse implicative. In *Proc. of Troisièmes Rencontres Internationales A.S.I.*, pages 339–349, Palermo,Italie, 2005.

[HGB05b] X. H. Huynh, F. Guillet, and H Briand. Arqat : An exploratory analysis tool for interstingness measures. In *Proc. of Applied Stochastic Models and Data Analysis*, pages 334–344, Brest,France, 2005.

[HGN00] J. Hipp, U. Güntzer, and G. Nakhaeizadeh. Algorithms for Association Rule Mining - a General Survey and Comparison. *SIGKDD Explorations*, 2:58–64, 2000.

[HPY00] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In Weidong Chen, Jeffrey Naughton, and Philip A. Bernstein, editors, *2000 ACM SIGMOD Intl. Conference on Management of Data*, pages 1–12. ACM Press, 05 2000.

[HYS05] T. Hamrouni, S. Ben Yahia, and Y. Slimani. Prince : An algorithm for generating rule bases without closure computations. In *Proc. of the 7th DaWaK Conference*, pages 346–355, 2005.

[Ise51] K. Iseki. On closure operation in lattice theory. In Nerd. Akad. Wetensch Proc. Ser. A 54, editor, *Idang. Math 13*, pages 318–320, 1951.

[Jac08] P. Jaccard. Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 44:223–270, 1908.

[KG00] M. Kryszkiewicz and M. Gajek. Concise representation of frequent pattern based on generalized disjonction-free generators. In *Proc. of Pacific-Asie Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 159–171, Kyoto,Japan, 2000.

[Kod99] Y. Kodratoff. Quelques contraintes symboliques sur le numérique en ecd et ect. In *SFDS*, pages 183–188, Grenoble,France, 1999.

[Ler81] I.C. Lerman. *Classification et analyse ordinale des donnés*. Dunod, 1981.

[Ler84] I. C. Lerman. Justification et validité statistique d'une échelle [0,1] de fréquence mathématique pour une structure de proximité sur un ensemble de variables observées, 1984. Rapport de recherche INRIA, Centre de Rennes IRISA.

[LFZ99] N. Lavrac, P. Flach, and B. Zupan. Rule evaluation measures : A unifying view. In G. Mineau and B. Ganter, editors, *Ninth International workshop on Inductive Logic Programming*, volume 1634, pages 174–185, 1999.

[LGR81] I.C. Lerman, R. Gras, and H. Rostam. Elaboration et évaluation d'un indice d'implication pour des données binaires. *Math Sc. Hum*, 74:5–35, 1981.

[LHCM00] B. Lui, W. Hsu, S. Chen, and Y. Ma. Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems*, 15(5):47–55, 2000.

[LK98] Dao-I Lin and Zvi M. Kedem. Pincer search: A new algorithm for discovering the maximum frequent set. *Lecture Notes in Computer Science*, 1377:105–121, 1998.

[LMV+04] P. Lenca, P. Meyer, B. Vaillant, P. Picouet, and S. Lallich. Evaluation et analyse multi-critère des mesures de qualité des règles d'association. *RNTI-E-1*, pages 219–246, 2004.

[LMVP03] P. Lenca, P. Meyer, B. Vaillant, and P. Picouet. Aide multicritères à la décision pour évaluer les indices de qualité de connaissances. In *Proc. of the EGC Conference*, volume 17, pages 271–282, Lyon,France, 2003.

[Loe47] J. Loevinger. A symetric approach to the construction and evaluation of tests of ability. *Psychological Monographs*, 61:1–49, 1947.

[LT04] S. Lallich and O. Teytaud. Evaluation et validation de mesures d'intérêt des règles d'association. *RNTI-E-1*, spécial:193–217, 2004.

[Lux91] M. Luxenburger. Implications partielles dans un contexte. *Math. Inf. Sci. hum.*, 113:35–55, 1991.

[Mon03] B. Monjardet. The presence of lattice theory in discrete problems of mathematical social sciences. why? *Mathematical Social Sciences*, 46:103–144, 2003.

[Mor62] J. Morgado. A characterization of the closure operator by mean of one axiom. *Portugal Math*, 21:155–156, 1962.

[MT96] H. Mannila and H. Toivonen. Multiple uses of frequent sets and condensed representation. In *Proc. of the KDD'96 Conference*, pages 189–194, Montreal,Canada, 1996.

[MT97] H. Mannila and H. Toivonen. Levelwise search and borders of theories in knowledge discovery. *Data Mining Knowledge Discovery*, 1. 3:241–258, November 1997.

[Nar82] H. Narushima. Principle of inclusion-exclusion on partially order set. *Discete Mathematics*, 42:243–250, 1982.

[Ö44] O Öre. Galois connections. *Transaction of the American Mathematical Society*, 55:494–513, 1944.

[OKO04] M. Ohsaki, S. Kitaguchi, and K. Okamoto. Evaluation of rules interstigness measures with a clinical dataset on hepatitis. In *ECML-PKDD, Pisa, Italy*, pages 362–373, 2004.

[Pas00a] N. Pasquier. *Data Mining : Algorithmes d'extraction et de réduction des règles d'associationdans les bases de données.* PhD thesis, Clermond-Ferrand II, Clermont-Ferrand, FRANCE, 2000.

[Pas00b] N. Pasquier. Extraction de bases pour les règles d'association à partir des itemsets fermés fréquents. In *Proc. 18èmes Congrés sur Informatique Organisations et Systèmes d'Information et de Décision*, pages 177–196, Var,France, 2000.

[PBTL99a] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rules using closed itemset lattices. *Information Systems*, 24:25–46, 1999.

[PBTL99b] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Closed set based discovery of small covers for association rules. In *Proc. 15emes Journees Bases de Donnees Avancées, BDA*, pages 361–381, Bordeaux,France, 1999.

[Pea88] J. Pearl. *Probabilistic reasoning in intelligent systems.* Morgan Kaufmann, 1988.

[PHM00] J. Pei, J. Han, and R. Mao. CLOSET: An efficient algorithm for mining frequent closed itemsets. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 21–30, Dallas,U.S.A., 2000.

[Plo73] C.R. Plott. Path independance, rationality and social choice. *Econometrica*, 41:1075–1091, 1973.

[PS91] G. Piatetsky-Shapiro. Discovery, analysis, and representation of strong rules. *Knowlegde Discovery in Databases*, AAAI Press/The MIT Press:229–248, 1991.

[SA96] R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. In *Proc. of 5th Biennial International Conference on Extending Database Technology (EDBT'96)*, volume 1057, pages 3–17, Avignon,France, 1996.

[SON95] A Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large databases. In *Proc. of the 21th VLDB Conference*, pages 432–444, Zurich,Switzerland, September 1995.

[SS88] M. Sebag and M. Shoenauer. Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases. In *Proc. of the European Knowledge Acquisition Workshop Conference*, pages 28–1–28–20, Bonn,Germany, 1988.

[STB+01] G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, and L. Lakhal. Intelligent structuring and reducing of association rules with formal concept analysis. In F. Baader, G. Brewka, and T. Eiter, editors, *Advances in Artificial Intelligence*, KI 2001, LNAI 2174, pages 335–350. Springer-Verlag, 2001.

[STB+02] G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, and L. Lakhal. Computing iceberg concept lattices with TITANIC. *Data and Knowledge Engineering*, 42:189–222, 2002.

[Toi94] H. Toivonen. Sampling large databases for association rules. In *Proc. of the 22nd VLDB Conference*, pages 134–145, San Diego,Chile, September 1994.

[Tot03] A. Totohasina. Normalisation de mesures probabilistes de la qualité des règles. In *Proc. SFDS'03, XXXV ième Journées de Statistiques*, volume 2, pages 985–988, Lyon, France, 2003.

[TR05] A. Totohasina and H. Ralambondrainy. Ion : a pertinent new measure for mining information from many types of data. In *IEEE SITIS'05.*, pages 202–207, Yaoundé, Cameroun, 2005.

[TRD04] A. Totohasina, H. Ralambondrainy, and J. Diatta. Notes sur les mesures probabilistes de la qualité des règles d'association : un algorithme efficace d'extraction des règles d'association implicatives. In *Proc. of CARI'04*, pages 511–518, Tunis, Tunisie, 2004.

[TRD05] A. Totohasina, H. Ralambondrainy, and J. Diatta. Une vision unificatrice des mesures probabilistes de la qualité des règles d'association booléennes et un algorithme efficace d'extraction des règles d'association implicatives. In *Proc. of TAIMA'05*, pages 375–380, Tunis, Tunisie, 2005.

[Wil82] R. Wille. Restructuring lattice theory: an approach based on hierarchies of concepts. In I. Rival, editor, *Ordered sets*, pages 445–470. Ridel, Dordrecht-Boston, 1982.

[WMSZ94] J. T. L. Wang, G. W. Marr, D. Shasha, and K. Zhang. Combinatorial pattern discovery for scientific data: Some preliminary results. In *Proc. of the ACM SIGMOD International Conference on Management of Data*, pages 115–125, Minneapolis,U.S.A., 1994.

[WZZ04] X. Wu, C. Zhang, and S. Zhang. Efficient mining of both positive and negative association rules. *ACM Transactions on information Systems*, 3:381–405, 2004.

[Zak00a] M. J. Zaki. Generating non-redundant association rules. In *Knowledge Discovery and Data Mining*, pages 34–43, 2000.

[Zak00b] M. J. Zaki. Scalable algorithms for association mining. In *IEEE Transactions on Knowledge and Data Engineering*, volume 12(3), pages 372–390, 2000.

[ZH99] M. J. Zaki and C.-J. Hsiao. CHARM: An efficient algorithm for closed itemset mining, 1999. Technical Report 99-10, Computer Science, Rensselaer Polytechnic Institute, 1999.

[Zha00] T. Zhang. Association rules. In *PAKDD 2000*, LNAI 1805, pages 245–256. Springer-Verlag, 2000.

[ZPOL97] M. J. Zaki, S. Parthasarathy, M Ogihara, and W. Li. New algorithms for discovery association rules. In *Knowledge Discovery and Data Mining*, pages 283–296, 1997.

RÉSUMÉ DE LA THÈSE

Les règles d'association révèlent des régularités non triviales et potentiellement utiles pour l'aide à la décision, dans des bases de données. Leur validité est évaluée par le biais de mesures de qualités dont les plus utilisées sont le support et la confiance. Pour une base de données transactionnelles d'un supermarché, elles sont du type « 90% des clients ayant acheté du vin et du fromage ont également acheté du pain, sachant que 75% des clients ont acheté ces articles».

Dans ce travail, nous spécifions une classe de mesures de qualité normalisées en ce sens qu'elles reflètent les situations de référence comme l'incompatibilité, la dépendance négative, l'indépendance, la dépendance positive, et l'implication logique entre la prémisse et le conséquent. Nous caractérisons les mesures de qualité normalisables, donnons les formules de normalisation et montrons que la plupart de celles qui sont normalisables ont la même normalisée, à savoir la mesure de qualité M_{GK} introduite dans Guillaume (2000). De plus, nous caractérisons des bases pour les règles positives et les règles négatives valides au sens de M_{GK}, et proposons des algorithmes de génération de ces bases.

Mots-clés : Base, Contexte binaire, Mesures qualité, Normalisation, Opérateur de fermeture, Règles d'association, Règles négatives, Relation binaire.

QUALITY MEASURES FOR ASSOCIATION RULES : NORMALIZATION AND CHARACTERIZATION OF BASES.

Summary

Association rules reveal non trivial patterns, potentialy useful for decision making, from huge databases. Their interestingness is assessed by means of quality measures, the most used of which being the support and the confidence. A standard example of an association rule for a transanctional database is of the form: « 90% of consumers who bought wine and cheese also bought bred, given 75% of consumers actually bought theses three items ».

In the present work, we specify a class of quality measures said to be normalized in the sense that their values reflect reference situations such as incompatibility, negative dependence, independence, positive dependence and logical implication between the premisse and the consequent of a rule. We characterise normalizable quality measures, give normalization formula and show that most of the normalizable quality measure dealt with in the literature have the same associated normalized one, namely the quality measure M_{GK} introduced in Guillaume (2000). We characterize bases for both positive and negative M_{GK}–valide association rules, and propose algorithms for generating these bases.

Key-words : Base, Binary context, Quality assessing, Normalization, Closure operator, Association rules, Negative rule, Binary relation.