Contributions à l'amélioration de génération des bases des règles d'association MGK-valides et applications en didactique des mathématiques

Harrimann Ramanantsoa

UNIVERSITÉ D'ANTANANARIVO
École doctorale : Problématiques de l'Éducation et Didactiques des Disciplines

THÈSE

Présentée à l'Université d'Antsiranana
Équipe d'accueil : Éducation et Didactiques des Mathématiques et de l'Informatique

Pour l'obtention du grade de

DOCTEUR DE L'UNIVERSITÉ D'ANTANANARIVO

Spécialité : **Didactiques des Mathématiques et de l'Informatique**
par

RAMANANTSOA Harrimann

Contributions à l'amélioration de génération des bases des règles d'association M_{GK}-valides et applications en didactique des mathématiques

---

Soutenue publiquement le 21 avril 2016 devant le jury composé de :

RAZAFIMBELO Judith	Professeur Titulaire	Université d'Antananarivo	Président
DIATTA Jean	Professeur Titulaire	Université de La Réunion	Rapporteur interne
HARISON Victor	Professeur Titulaire	Université d'Antananarivo	Rapporteur externe
RAKOTOSON Jean Emile	Professeur	Université de Fianarantsoa	Examinateur
FENO Daniel Rajaonasy	Maître de conférences	Université de Toamasina	Examinateur
TOURNÈS Dominique	Professeur Titulaire	Université de La Réunion	Co-Directeur
TOTOHASINA André	Professeur Titulaire	Université d'Antsiranana	Directeur

Remerciements

Cette thèse a été préparée conjointement au laboratoire de mathématiques et informatique de l'École Normale Supérieure pour l'Enseignement Technique (ENSET), université d'Antsiranana et au Laboratoire d'Informatique et de Mathématiques (LIM), université de La Réunion, dans le cadre de la recherche multidimensionnelle touchant les thèmes de fouille des données et de didactique des mathématiques.

Que l'Agence Universitaire de la Francophonie, Bureau Océan Indien (AUF-BOI), trouve ici, l'expression de mes remerciements les plus sincères pour ses soutiens financiers, dans le cadre du projet Horizons Francophones Sciences Fondamentales (HF-SF) me permettant, entre autres, d'effectuer des mobilités scientifiques durant les trois années de préparation de thèse (séjours au laboratoire LIM de l'université de La Réunion, regroupements à Antananarivo et à l'université de Yaoundé, Cameroun) et de faire venir mes deux membres du jury provenant de l'université de La Réunion.

Je tiens à exprimer mes plus vifs remerciements à Monsieur TOTOHASINA André, Professeur Titulaire à l'université d'Antsiranana, Madagascar, Responsable scientifique de l'équipe d'accueil Éducation et Didactique de Mathématiques et de l'Informatique (EDMI) et à Monsieur TOURNÈS Dominique, Professeur à l'université de La Réunion, France, Directeur de L'Institut de Recherche sur l'Enseignement de Mathématiques (IREM), pour m'avoir proposé le sujet de recherche. Leurs conseils, leurs rigueurs scientifiques et leurs encouragements m'ont permis de franchir bien des obstacles. Qu'ils trouvent ici mes plus grands respects et l'expression de mes reconnaissances les plus sincères.

Je tiens également à adresser mes vifs remerciements ainsi que ma profonde gratitude à Monsieur DIATTA Jean, Professeur à l'université de La Réunion, France, et Directeur du LIM qui m'a toujours réservé un accueil chaleureux lors de mes nombreux séjours dans son laboratoire et qui a aussi accepté d'être le rapporteur interne de mes travaux.

Que Monsieur HARISON Victor, Professeur Titulaire à l'Université d'Antananarivo, Madagascar, Directeur général de l'Institut national des Sciences Comptables et de l'Administration d'Entreprises (INSCAE), trouve ici l'expression de mes remerciements les plus sincères d'avoir accepté d'être le rapporteur externe malgré son emploi du temps très chargé.

Je remercie également Madame RAZAFIMBELO Judith, Professeur Titulaire à l'Université d'Antananarivo, Madagascar, Directrice de l'École Doctorale Problématique de l'Éducation et Didactique de Disciplines (PE2Di) de m'avoir inscrit parmi ses doctorants et pour l'honneur qu'elle me fait d'avoir acceptée d'être le Président du jury.

Ma gratitude va aussi à l'ensemble des membres du jury. Je vous adresse mes remerciements les plus respectueux.

Je profite de cette occasion pour remercier l'ensemble des collègues enseignants de l'École Normale Supérieure pour l'Enseignement Technique (ENSET) et de l'institution Saint Joseph Antsiranana de m'avoir soutenu et de couvrir mes multiples absences durant toutes les années de préparation de cette thèse. Sans oublier les amis, les étudiants Malagasy à l'université de La Réunion, les collègues doctorants comme BEMARISIKA Parfait, RANDRIANTSIFERANA Rivo Sitraka et bien d'autres encore, ils ont su créer des atmosphères favorables et agréables au milieu des durs labeurs. Ils ont su faire de ces quatre années des moments inoubliables.

Enfin, mes reconnaissances les plus sincères vont à l'endroit de mon épouse et mes deux filles ainsi qu'à l'ensemble de ma famille, leur soutien moral et affectif inconditionnel ont fait de moi ce que je suis aujourd'hui.

Merci à tous !!

Sommaire

Liste des tableaux

Liste des figures

Liste des abréviations

ASI	:	Analyse Statistique Implicative
BC	:	Base de Couverture
BDG	:	Base de Duquenne-Guigues
BG	:	Base Générique
BI	:	Base Informative
BNA	:	Base Négative Approximative
BNE	:	Base Négative Exacte
BP	:	Base Propre
BPA	:	Base Positive Approximative
BPE	:	Base Positive Exacte
CHIC	:	Classification Hiérarchique Implicative et Cohésitive
$CPIR$	:	Conditional Probability Incrementation Ratio
EIAH	:	Environnement Informatique d'Apprentissage Humain
FF	:	Fermé Fréquent
ION	:	Implication Statistique Orientée Normée
M_{GK}	:	Mesure de Guillaume-Khenchaff
$NBNAD$	:	Nouvelle Base Négative Approximative à Droite
$NBNE$	:	Nouvelle Base Négative Exacte
$NBPA$	:	Nouvelle Base Positive Approximative
$NBPE$	:	Nouvelle Base Positive Exacte
PFF	:	Pseudo-fermé Fréquent
RNA	:	Règle Négative Approximative
RNE	:	Règle Négative Exacte
RPA	:	Règle Positive Approximative
RPE	:	Règle Positive Exacte
RTI	:	Réduction Transitive de la base Informative
$SBNA$	:	Sémi-Base Négative Approximative
$SBPA$	:	Sémi-Base Positive Approximative
$SBPE$	:	Sémi-Base Positive Exacte

Chapitre 1

Introduction générale

1.1 Contexte général

Au moment où, au niveau mondial, l'intégration pédagogique des Technologies de l'Information et de la Communication (TIC) est devenue un important sujet de préoccupation dans la sphère de l'éducation, dans le contexte éducatif Malagasy, les acteurs de l'éducation s'inquiètent de la désaffection des élèves envers les disciplines scientifiques, notamment, les mathématiques. Ce phénomène peut être ressenti à travers les statistiques portant sur les candidats au baccalauréat. Sur l'ensemble des candidats, le pourcentage des candidats inscrits dans la série scientifique (série C) n'a pas dépassé la barre de 6% durant ces dix dernières années (statistiques publiées sur le site du ministère de l'Enseignement supérieur) ; pour les sessions 2014 et 2015, ce pourcentage est descendu à 4%. Il semblerait que la société environnante, y compris les parents, encourage une telle démotivation. Un rapport de recherche d'une équipe qui travaille sous la couverture de l'United Nations Educational, Scientific and Cultural Organization (UNESCO), dirigée par le Professeur Michèle Artigue (Artigue, 2011) montre qu'il n'y a pas qu'à Madagascar qu'on se demande « comment redynamiser l'enseignement des mathématiques ? » Selon ce rapport, bien que notre société et nos environnements entretiennent des liens très étroits avec la science et la technologie, donc avec les mathématiques, plusieurs défis restent à relever pour qu'un enseignement pertinent et de qualité de ces dernières puisse être, d'une part, accessible à tous et, d'autre part, amener les apprenants à développer des réflexions critiques et de la créativité dans le but de former des citoyens capables de prendre part aux problèmes que le monde doit affronter (sur l'énergie, l'environnement, la santé, etc.). Le même rapport affirme que « *l'enseignement des mathématiques dans la scolarité de base est trop souvent encore un enseignement peu stimulant : dans lequel les pratiques expérimentales, les activités de modélisation sont rares* (...) » et « *penser une éducation de qualité pour tous aujourd'hui ne peut se faire sans prendre en compte la dimension technologique* ». Par rapport à ces réflexions, les TIC peuvent être utilisées pour stimuler, et par la suite, améliorer l'enseignement et l'apprentissage des mathématiques. Dans la littérature, beaucoup d'auteurs affirment que les TICE sont des bons outils de conjecture dans la mesure où elles permettent de réduire les coûts des essais-erreurs (Emprin, 2008 ; Artigue, 2008 ; BRAGA, 2009 ; Ramanantsoa *et al.*, 2012 ; Bemarisika *et al.*, 2012). Pour le cas du système éducatif Malagasy, l'intégration pédagogique des TIC peut être considérée comme une solution face aux problématiques sur la désaffection des élèves à l'égard des disciplines scientifiques, en particulier pour les mathématiques. Étant donnés les paradoxes rapportés dans (Emprin, 2008), entre la faiblesse, la non-diversité de l'utilisation

des TICE[1] et les efforts institutionnels consentis dans l'utilisation des outils informatiques en éducation dans plusieurs pays d'Europe, la question se pose : « comment faire ? » Comment introduire en classe ces nouveaux outils d'enseignement et d'apprentissage pour que ces derniers puissent être efficaces et efficients, compte tenu des multiples contraintes qu'on doit prendre en compte ? Pour essayer d'apporter des éléments de réponse à ces questions, nous allons faire un tour d'horizon sur les résultats d'études faites ailleurs et surtout en Afrique en matière d'intégration pédagogique des TIC.

1.2 Quelques problèmes sur l'intégration pédagogique des TIC

Dans un autre rapport de l'UNESCO, publié en 2004, on peut lire les informations suivantes « *L'intégration des ordinateurs et des technologies dans les écoles est un processus coûteux et parfois complexe, qui requiert toute une série d'équipements, un personnel compétent pour l'installation et le fonctionnement, un support technique et une formation des autres utilisateurs au bon usage de ces matériels. Mais les avantages évidents qu'elle apporte aux écoles et à leurs élèves sont assez significatifs pour que l'introduction des technologies dans les classes soit désormais l'une des priorités des planificateurs de l'éducation, aussi bien dans les pays développés que dans le pays en développement, même si les défis et les obstacles à surmonter sont souvent très différents (Pelgrum et Law, 2004).* » Dans ces informations, on peut retenir que quel que soit le développement d'un pays, introduire des outils informatiques dans les pratiques pédagogiques est très souvent difficile, et les obstacles et difficultés peuvent changer d'un pays à l'autre. Ceci dit, dans chaque pays, on doit chercher des approches adaptées au contexte local. D'où l'importance du volet de recherche-action dans le processus d'intégration pédagogique de TIC.

Puisque le système éducatif Malagasy a plusieurs facteurs communs avec ses voisins Africains, nous allons voir ce que ces derniers ont fait. Un projet panafricain a été mis en place par une équipe du Professeur Thierry Karsenti dans le but de : « *mieux comprendre comment l'intégration pédagogique des TIC peut améliorer la qualité des enseignements et des apprentissages en Afrique.* » En 2009, après avoir observé une centaine de salles de classes dans plus de 15 pays d'Afrique (Karsenti, 2009), ils ont publié un rapport qui va nous permettre d'avoir une idée sur la situation d'intégration pédagogique des TIC en Afrique. Dans ces études, on a établi quatre niveaux d'intégration pédagogique des TIC :

A : Les technologies sont considérées comme objet d'apprentissage et l'enseignement est centré sur le professeur.

B : Les technologies sont toujours considérées comme objet d'apprentissage, mais l'enseignement est plutôt centré sur l'élève.

Le niveau A se résume à un cours magistral d'informatique et le niveau B, c'est toujours un cours d'informatique, mais il y a beaucoup plus de participation des élèves. Ce sont peut-être des étapes nécessaires, mais on ne doit pas en rester là parce qu'ici les élèves ne bénéficient pas des apports des technologies dans l'apprentissage des disciplines spécifiques.

C : À ce niveau, les enseignants utilisent les technologies dans des diverses disciplines. Les technologies ne sont plus des objets d'apprentissage, mais utilisées comme outils

1. TICE Technologie de l'Information et de Communication pour l'Enseignement

pour l'enseignant. Par rapport aux deux premiers niveaux, celui-ci est un réel début d'intégration des TIC dans l'enseignement.

D : Les technologies sont utilisées pour amener les élèves à s'approprier des connaissances. À ce niveau, ce sont les élèves qui sont appelés à faire usage des TIC pour apprendre les diverses disciplines.

Par rapport à l'ensemble des établissements observés, 80% restent coincés au niveau A et B (en 2009) et un peu moins de 12% se trouvent au niveau C. Ces études montrent un bilan sur l'intégration pédagogique des TIC en Afrique aux alentours de l'année 2009. Elles soulignent qu'au début du processus, la tendance est de considérer les technologies comme objets d'apprentissage en laissant de côté les autres disciplines.

Vu ce qui se passe dans des établissements Malagasy où on initie les élèves aux cours d'informatique, il est clair qu'une bonne partie des établissements restent toujours aux niveaux A et B. De plus, du moment que les enseignants en poste n'arrivent pas à associer les nouveaux outils à leurs disciplines, ils risquent de percevoir l'intégration des TICE comme un exercice indépendant de leurs disciplines, donc des tâches supplémentaires qui viennent déséquilibrer leurs méthodes de travail. Donc, ils feront tout pour résister à l'intégration des nouveaux outils.

Par rapport à ce constat, les chercheurs ont des rôles à jouer pour convaincre les enseignants, les dirigeants que l'intégration pédagogique des TIC va au-delà des niveaux A et B ; il ne faut donc pas rester au stade d'alphabétisation numérique (niveau A et B). On peut se demander maintenant « comment montrer aux gens (enseignants, décideurs, etc.) les apports des TIC dans l'enseignement de leurs disciplines ? Comment les convaincre pour changer de pratique pédagogique ? »

D'autres études faites en Afrique par une équipe des chercheurs dans le domaine de l'intégration pédagogique des TIC sont rapportées dans (Karsenti *et al.*, 2012). Elles mettent en avant les informations ci-après. L'absence de politiques d'intégration des TIC clairement définies, l'absence de financement pour l'acquisition des matériels, ratios élève/ordinateur trop élevés, instabilité des courants électriques, vétusté des appareils, etc., constituent des facteurs communs aux obstacles rencontrés dans divers pays d'Afrique. Dans la majorité des pays observés, les enseignants ne sont pas formés à l'utilisation pédagogique des TIC et les programmes destinés aux futurs enseignants offrent peu d'occasion d'apprendre les habiletés nécessaires pour intégrer les TIC dans l'enseignement. Plusieurs arguments peuvent être utilisés pour affirmer que les situations à Madagascar sont assez similaires à celles qui viennent d'être citées. Encore une fois, si on veut inverser la tendance actuelle, il faudra convaincre les forces vives de l'éducation à prendre toutes les mesures nécessaires pour que les élèves puissent tirer profit de ces technologies. Un des moyens pour en convaincre les gens est de leur « montrer » l'efficacité de l'approche adoptée. En éducation, en particulier en didactique des mathématiques, cela est possible grâce à des expérimentations et à l'analyse des données issues de ces expérimentations.

Donc, pour redynamiser l'enseignement des mathématiques, on peut faire appel aux TICE et, pour que l'intégration pédagogique des TIC puisse être effective, il faut inclure non seulement les enseignants, mais aussi toutes les forces vives qui peuvent apporter des parts de briques à l'amélioration de la qualité de l'éducation. Il faut donc convaincre les acteurs de l'éducation, ceci par le biais d'une expérimentation suivie de l'analyse des données. Pour valider une conjecture, les chercheurs en didactique des mathématiques ont souvent recours

à cette pratique d'expérimentation suivie de l'analyse des données. Nous pouvons donner des exemples qui confortent cette affirmation sur la nécessité d'observation et d'expérimentation. La totalité des rapports cités jusqu'à présent sont basés sur des expérimentations ou des observations suivies des analyses des données. Ces arguments nous amènent droit à l'un des objectifs de cette thèse : « *amélioration d'un outil statistique qui peut être utilisé pour analyser des données issues d'une expérimentation didactique* ».
Selon l'enchainement des idées que nous venons de présenter, l'analyse des données a une place importante dans ce processus très complexe de l'intégration pédagogique des TIC et donc, de l'amélioration de la qualité d'enseignement et d'apprentissage des mathématiques. Afin de minimiser les risques d'erreur dans les résultats d'analyse et donc dans des éventuelles décisions que l'on peut prendre, il est impératif que l'outil d'analyse des données utilisé soit fiable. Cela nous amène au paragraphe consacré aux outils statistiques que l'on peut utiliser efficacement pour analyser des données d'une expérimentation ou d'une observation didactique, entre autres.

1.3 Outils d'analyse des données utilisés pour la recherche en didactique

Certes, tous les outils statistiques utilisés dans d'autres domaines peuvent servir pour analyser des données issues des expérimentations didactiques et inversement. Autrement dit, les outils statistiques sont fondamentalement conçus indépendamment du domaine d'utilisation, ils dépendent seulement des types des données utilisées et des objectifs de l'analyse.
Souvent, les chercheurs en didactique des disciplines, même en didactique des mathématiques, ne sont pas forcement des statisticiens. Par conséquent, parmi la multitude des outils mathématiques et informatiques (i. e. logiciels) d'analyse des données qui sont disponibles, les chercheurs en didactique se contentent d'utiliser les « plus simples », à savoir l'analyse statistique descriptive, unidimensionnelle dans la plupart du temps, c'est à dire, en étudiant la description (via les paramètres des positions, des pourcentages, des dispersions et des concentrations) de chacune des variables étudiées indépendamment des autres et très rarement en multidimensionnelle avec quoi on peut faire la description en tenant compte des interactions entre les variables. Il est bien clair qu'avec l'analyse statistique descriptive, on peut déjà obtenir plusieurs informations pertinentes concernant la situation étudiée. Mais souvent, on peut se demander s'il y a des corrélations entre les observations. Mieux encore, on peut se demander s'il y aurait (dans les données) des relations de co-occurence, et comment les découvrir objectivement ? L'extraction des règles d'association est un outil statistique permettant d'extraire et de valider des relations complexes entre les observations (Cadot, 2006). En 1979, Régis GRAS a initié l'utilisation des outils statistiques permettant de mesurer l'intensité d'implication entre des variables binaires pour analyser des données issues des observations ou expérimentations didactiques. Au début des années 90, l'équipe de recherche dirigée par Régis GRAS a conçu l'outil CHIC pour fouiller les règles d'association implicative et statistique de type « si variable X, alors variable Y » (Gras et Larher, 1992 ; Gras, 1996). Des fondements théoriques de ces outils statistiques sont donnés dans (Gras et Régnier, 2009). En 1993, Agrawal *et al.* ont introduit le concept d'extraction des règles d'association fondé sur les deux mesures de qualité « Support et Confiance ». Depuis, les théories ne cessent de foisonner et d'évoluer. Actuellement, on compte plusieurs dizaines de

mesures. Selon la théorie sur l'extraction des règles d'association, la qualité des règles extraites dépend fortement de la mesure utilisée. Grâce à l'évolution de l'informatique, le coût de stockage des données diminue de plus en plus et il est devenu plus facile de manipuler des données à plusieurs dizaines, voire des centaines de variables. Le revers de la médaille est que l'utilisateur (l'analyste des données par exemple) peut se retrouver en face d'un gros volume d'information à interpréter. Autrement dit, l'outil d'extraction des connaissances à partir des données peut fournir un nombre trop important d'informations dont certaines sont redondantes jusqu'au point où l'utilisateur se retrouve au point de départ : Extraire des informations à partir des données, seulement, cette fois, les données sont les résultats de l'analyse des données de départ. Ce problème a conduit les chercheurs dans ce domaine à concevoir un nouveau concept qui s'appelle « bases des règles d'association ». Une base des règles est un ensemble minimal des règles à partir duquel on peut déduire les autres règles en utilisant ce qu'on appelle « axiomes d'inférences ». Depuis, plusieurs questions tournant autour du concept des bases des règles ont été posées et étudiées.
D'autres part, sur le plan éducatif à Madagascar, plusieurs problématiques et questionnements ont été posés. En commençant par la désaffection des élèves à l'égard des disciplines scientifiques, en particulier les mathématiques ; l'intégration pédagogique des TIC peut constituer un élément de réponse à ce problème. Pourtant, cette dernière ne peut être effective que si tous les acteurs y prennent part ; donc, il faut les convaincre. D'où la nécessité d'expérimentation et d'analyse des données récoltées. Or la fiabilité des résultats dépend de la qualité de l'outil statistique utilisé. Par rapport à tout cela, nous allons présenter les problématiques de nos travaux.

1.4 Objectifs et organisation de la thèse

Pour des raisons que nous développerons au troisième chapitre du présent rapport, nous avons choisi de travailler avec la mesure M_{GK} qui est parfois dénommée ION : Implication Orientée et Normalisée (Totohasina *et al.*, 2004 ; Totohasina et Ralambondrainy, 2005) ou $CPIR$: Conditional Probability Incrementation Ratio (Wu *et al.*, 2004). Des études ont été déjà faites concernant cette mesure. Nous avons constaté que les descriptions et les algorithmes d'extraction des bases des règles valides selon le couple de mesures $Support - \mathrm{M}_{GK}$ représentent un certain nombre des limites qu'on pourrait améliorer, malgré son grand avantage par rapport aux autres. À partir de ces constats, nous nous sommes fixés comme objectifs de trouver des moyens pour améliorer les descriptions et les algorithmes d'extraction des bases des règles valides selon les mesures $Support$ et M_{GK}, suite aux travaux de Feno (2007). Les questions auxquelles nous essayons de répondre sont les suivantes :

- Est-il possible d'améliorer la qualité des règles dans les bases ?
- Est-il possible de diminuer encore une fois le nombre des règles dans les bases ? Si oui, pourquoi et comment faire ?
- Comment améliorer les algorithmes d'extraction des bases des règles ?
- Enfin, comment utiliser ces outils en faveur de la recherche en didactique des mathématiques ?

Pour essayer d'apporter des éléments de réponse à ces questions, nous avons structuré ce rapport en deux grandes parties. La première, précédée par ce chapitre d'introduction décrivant les contextes généraux et les objectifs de la thèse, est constituée par deux chapitres contenant des états de l'art sur les outils d'Analyse Statistique Implicative (ASI) et ses ap-

plications en didactique des mathématiques suivies des exemples sur des données concrètes. Dans cette partie, nous expliciterons le pourquoi de notre choix de travailler avec la mesure M_{GK} et aussi les états de l'art sur le concept des bases des règles d'association. La deuxième partie constituée de deux chapitres rapporte nos contributions. Nous commencerons par la redéfinition des bases des règles M_{GK}-valides en essayant, dans la mesure du possible d'améliorer les limites constatées dans la première partie. Toujours dans l'objectif d'optimiser le nombre des règles dans les bases, nous proposerons un nouveau concept que nous avons appelé « semi-base des règles ». Nous proposerons ensuite des algorithmes d'extraction de ces nouvelles bases des règles. Enfin, nous terminerons cette deuxième partie par un chapitre consacré à l'application de ces nouveaux outils statistiques dans l'analyse des données issues de nos expérimentations. Cette deuxième partie sera suivie d'un chapitre de conclusions et des perspectives.

Première partie

Première contribution en didactique et état de l'art sur les bases des règles d'association

Chapitre 2

Découverte des liens implicatifs et ses applications en didactique

Sommaire

2.1 Introduction

Connaître les difficultés et les origines des erreurs de ses élèves, avoir des idées sur les éventuels effets d'une nouvelle approche ou nouveaux outils pédagogiques constituent des préoccupations quotidiennes de tout enseignant professionnel dans l'exercice de sa fonction. Pour atteindre ses objectifs, un enseignant a souvent recours aux tests ou évaluations, aux expérimentations. Dans un premier temps, ces derniers peuvent aboutir sur un volume de données

binaires. Dans le cas particulier de l'enseignement, ces données peuvent être des productions d'élèves, des réponses à un questionnaire, etc. Afin d'extraire des connaissances utiles dans un tel volume des données, ces dernières doivent être analysées et, en fonction de l'objectif de l'expérimentation et des types des données recueillies, on peut choisir les outils statistiques adéquats. Dans notre cas, nous allons nous intéresser à la découverte des liens de cause à effet entre les variables observées. En effet, les analyses statistiques descriptives des données d'expérimentation telle que l'étude des moyennes, des pourcentages, des valeurs minimales et maximales, permettent déjà de mettre en évidence quelques comportements de la population étudiée (une classe, ensemble des classes, ensemble des travaux d'un élève particulier, etc.). Mais ce type d'analyse, même si son utilisation est relativement généralisée, en particulier dans la sphère de l'éducation, ne permet pas de mettre en lumière les éventuels liens entre les observations. Des liens qui peuvent être très utiles pour étudier les causes de certaines observations. Avec l'analyse descriptive, on peut par exemple connaitre la proportion des élèves qui ont commis telle ou telle erreur à partir des seules données d'observation ; mais par contre, si les erreurs sont liées, on ne peut pas savoir quelles sont les causes et quelles sont les conséquences. Pour dépasser cette limite, et découvrir des liens possibles dans les observations, on peut faire appel à l'extraction des règles d'association. Une analyse statistique permettant de découvrir des liens de co-occurrences de type : « à chaque fois qu'un groupe des variables (motif) X est présent chez un individu, alors un autre groupe des variables Y est très souvent présent chez cet même individu ». Relativement à un contexte dans lequel on a considéré un ensemble des variables qui peuvent former un « tout » vis-à-vis d'un objet d'étude bien précis, ces liens implicatifs (non symétrique) de co-occurences peuvent être interpreter comme des liens de causalité de type « *si A, alors B* » que l'on note $A \rightarrow B$. On peut par exemple établir des liens entre les erreurs des élèves, leurs compétences, entre les erreurs et les outils pédagogiques utilisés. La question qui se pose maintenant est : « Comment valider objectivement ce type d'information ? ». Plusieurs outils mathématiques ont été proposés au fil des années pour essayer de découvrir et valider objectivement ce type de lien implicatif ou ce type d'association à partir d'un volume de données. Dans ce chapitre, nous allons voir quelques outils mathématiques sur lesquels sont fondés le concept de découverte des liens implicatifs à partir des données. Nous commencerons par présenter quelques généralités sur le concept d'extraction des règles d'association. Nous présenterons ensuite des propriétés de quelques mesures de qualité en commençant par détailler l'approche choisie par l'équipe de Gras *et al.*, suivi de la présentation d'un exemple d'étude sur l'effet de l'utilisation d'un vidéoprojecteur lors de l'enseignement de fonction numérique d'une variable réelle en classe de seconde. Enfin, après avoir présenté d'autres mesures de qualité, nous allons faire une analyse critique de l'utilisation du couple de mesures Support-Confiance et justifier ainsi les raisons qui nous ont poussé à utiliser ladite mesure M_{GK}.

2.2 Généralités sur l'extraction des règles d'association

Considérons une population $\mathcal{O} = \{o_1, o_2, ..., o_n\}$ de n objets (ou n individus), un ensemble $\mathcal{I} = \{i_1, i_2, ..., i_m\}$ de m variables (ou items) binaires et une relation binaire $\mathcal{R}$ de $\mathcal{O}$ vers $\mathcal{I}$. Pour chacune des variables binaires i_p l'observation consiste à déterminer si elle est présente ou non chez chacun des objets o_k de la population étudiée. La dénomination binaire vient du fait que pour un objet o_k, une variable i_p ne peut prendre que deux valeurs exclusives : elle prend la valeur 1 si le comportement qu'elle désigne est présent chez l'objet o_k (i. e. $i_p(o_k) = 1$),

dans ce cas, on dit que o_k est en relation avec la variable i_p et on note $o_k \mathcal{R} i_p$, elle prend la valeur 0 sinon (i. e. $i_p(o_k) = 0$). En guise d'exemple, on peut imaginer que les variables sont constituées d'un ensemble d'erreurs qu'un élève peut commettre ou d'un ensemble de démarches qu'il peut suivre, et une classe entière représente la population étudiée. Dans ce cas, l'expérimentation consiste à observer si une erreur i_p a été commise par un élève o_k et ceci pour tous les couples (o_k, i_p) de $\mathcal{O} \times \mathcal{I}$. Si on considère deux parties disjointes X et Y de $\mathscr{P}(\mathcal{I})$, l'extraction des règles consiste à déterminer si une règle de type « *Si* X*, alors* Y » que l'on note par « $X \rightarrow Y$ » est exactement valide ou valide avec une certaine marge d'erreur acceptable ou tout simplement non valide. Toujours dans notre précédent exemple, à l'issue des étapes d'extraction, si une règle $X \rightarrow Y$ est valide, alors elle pourrait s'interpréter de la façon suivante : « *c'est le groupe (partie de* $\mathcal{I}$*) d'erreur* X *qui est la cause du groupe d'erreur* Y ». Pour valider objectivement une règle d'association, il faut faire appel aux mesures de qualité. Relativement à une mesure de qualité μ fixée, une règle $X \rightarrow Y$ est dite valide lorsque sa mesure $\mu(X \rightarrow Y)$ dépasse un certain seuil calculé à partir d'un paramètre ou fixé l'utilisateur. Avant de découvrir et exploiter quelques mesures de qualité, nous allons d'abord présenter des définitions et terminologies qui vont être utiles dans la suite.

2.3 Quelques définitions et terminologies

Définition 2.1 (Contexte d'extraction)**.** *Étant donné un ensemble* $\mathcal{O}$ *des objets (des objets ou des transactions) et un ensemble* $\mathcal{I}$ *d'item (attribut ou variable), et une relation binaire* $\mathcal{R}$*, le triplet* $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ *forme ce qu'on appelle un contexte binaire d'extraction ou contexte. Un couple* (o, i) *de* $\mathcal{O} \times \mathcal{I}$ *appartient au graphe de* $\mathcal{R}$ *signifie que la transaction ou l'objet* o *contient l'item* i *et on note* $i(o) = 1$*. Pour le reste du document, on assimile* $\mathcal{R}$ *à son graphe.*

Exemple 1 (Exemple d'un contexte)**.**
Considérons l'ensemble d'objet $\mathcal{O} = \{o_1, o_2, o_3, o_4, o_5, o_6\}$*, l'ensemble d'item* $\mathcal{I}$ *tel que :* $\mathcal{I} = \{A, B, C, D, E, F\}$ *et* $\mathcal{R}$ *la relation binaire définie dans le tableau 2.1.*

	A	B	C	D	E	F
o_1	1	1	1	0	1	0
o_2	1	1	1	1	0	0
o_3	1	0	0	1	0	0
o_4	0	1	0	1	1	0
o_5	1	1	1	0	1	1
o_6	1	1	0	1	1	1

Tableau 2.1 – Contexte $\mathcal{K}$

Le triplet $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ forme un contexte d'extraction.

Définition 2.2.
Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ *un contexte d'extraction. Une partie* X *de* $\mathcal{I}$ *est appelée motif positif (ou itemset positif). Un motif* X *est présent chez un objet* o *si pour tout item* i *de* X*,* o *contient* i *(conjonction de présence).*

Formellement, pour tout $X \subset \mathcal{I}$ *et pour tout* o *de* $\mathcal{O}$, $X(o) = 1 \Leftrightarrow \forall i \in X, (o,i) \in \mathcal{R}$.
Par contraposition, on a l'équivalence : $X(o) = 0$ *si et seulement s'il existe* i *dans* X *tel que* $(o,i) \notin \mathcal{R}$ *(disjonction des absences). Dans ce cas, on dit que le motif* X *est absent de l'objet* o *; on note* $\overline{X}(o) = 1$ *et* $\overline{X}$ *est appelé motif négatif.*

Définition 2.3 (Extension et Intension)**.**
Considérons un contexte d'extraction $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ *et notons par* $\mathscr{P}(E)$ *l'ensemble des parties d'un ensemble quelconque* E.

a. *Pour tout* $X \in \mathscr{P}(\mathcal{I})$, *l'ensemble* X' *défini par :* $X' = \{o \in \mathcal{O} / \forall i \in X, (o,i) \in \mathcal{R}\}$ *est appelé extension de* X.

b. *Pour tout* $O \in \mathscr{P}(\mathcal{O})$, *l'ensemble* O' *défini par :* $O' = \{i \in \mathcal{I} / \forall o \in O, (o,i) \in \mathcal{R}\}$ *est appelé intension de* O.

c. *Pour tout motif* $X \in \mathscr{P}(\mathcal{I})$, *le Support de* X *est défini par :* $Supp(X) = \dfrac{Card(X')}{Card(\mathcal{O})}$ *et* $Supp(\overline{X}) = 1 - Supp(X)$

Propriété 2.1 (Motif négatif)**.** *Pour tout motif* X *d'un contexte* $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$, *l'extension d'un motif négatif* $\overline{X}$ *est égal au complémentaire de* X' *dans* $\mathcal{O}$ *:* $\overline{X}' = \mathcal{O} \setminus X'$. *L'action de négation est involutive :* $\overline{\overline{X}} = X$.

Exemple 2. *Considérons le contexte* $\mathcal{K}$ *du tableau 2.1, posons* $X = AB$[1] *et* $O = o_5o_6$.
L'extension de X *est :* $X' = o_1o_2o_5o_6$ *et* $Supp(X) = \frac{Card(X')}{Card(\mathcal{O})} = \frac{4}{6} = 0,66$.
L'intension de O *est :* $O' = ABEF$
Pour le motif négatif $\overline{X}'$, *on a :* $\overline{AB}' = o_3o_4$.

Terminologie

Pour toute règle $X \to Y$, le motif X est appelé la prémisse et Y est appelé le conséquent.

2.4 ASI selon l'approche de R. Gras (1979)

Partons toujours d'un contexte $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ et considérons deux parties disjointes X et Y de $\mathcal{I}$. Désignons par X' et Y' deux parties de la population $\mathcal{O}$ qui sont respectivement en relation avec les motifs X et Y : $X' = \{o \in \mathcal{O} / \forall\, i \in X, o(i) = 1\}$. D'une autre manière, X' représente l'ensemble des objets sur lesquels on a identifié la présence des comportements désignés par toutes les variables constituant le motif X ; même définition pour Y'. Pour mieux comprendre l'approche de R. Gras, on peut distinguer deux cas fondamentaux : implications exactes et implications approximatives ou quasi-implications.

2.4.1 Implication exacte

Supposons que $X' \subset Y'$ (voir Fig. 2.1 à gauche). Dans cette situation, tous les objets présentant les comportements désignés par le motif X (objets présentant simultanément les comportements désignés par toutes les variables qui composent le motif X) possèdent aussi

1. Pour simplifier la notation, un ensemble $\{A, B\}$ sera noté par AB.

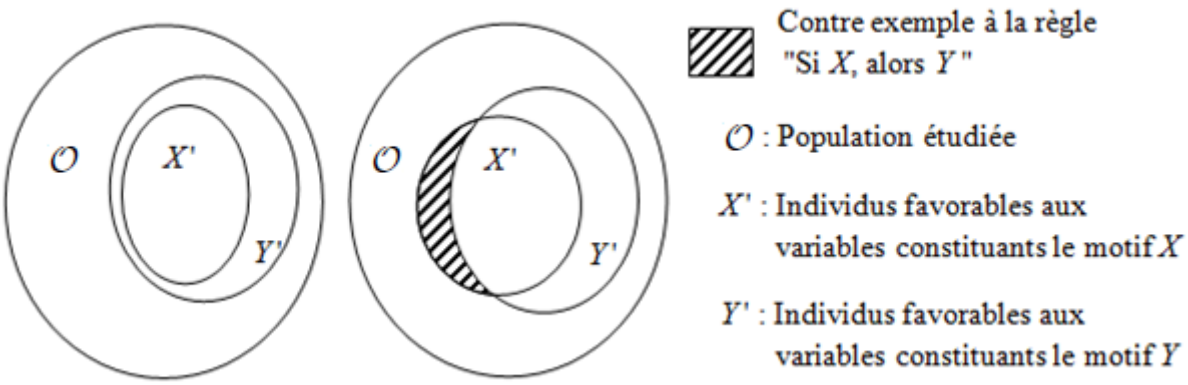


FIGURE 2.1 – Implication exacte (à gauche) et approximative (à droite)

les comportements désignés par le motif Y. En regardant dans l'ensemble des populations étudiées, on peut affirmer avec exactitude que pour un objet choisi au hasard, « *si cet objet possède les comportements désignés par le motif* X*, alors il possède aussi les comportements désignés par le motif* Y ». On a donc une règle exacte de type « si X, alors Y ». Si par exemple, X et Y représentent deux groupes d'erreurs possibles que chaque individu (objet) de la population étudiée peut commettre, la règle « si X, alors Y » signifie qu'à chaque fois qu'on observe le groupe d'erreurs X chez un objet (en l'occurrence un élève), on observe aussi le groupe d'erreurs Y chez le même objet. Ce type d'information pourrait être utile pour comprendre la cause des erreurs des élèves. En effet, on peut avoir ce type d'information lorsque le groupe d'erreurs Y était juste formé de conséquences du groupe d'erreurs X. Dans ce cas, ce n'est peut-être pas le groupe d'erreurs Y qu'il faut corriger (du moins dans un premier temps) mais plutôt le groupe d'erreurs X. Autrement dit, avant de « corriger Y », il faut d'abord s'attarder sur X, parce que Y est juste une conséquence de X. Cette information peut guider un enseignant ou un tuteur artificiel dans la pratique des travaux de remédiation. On peut aussi imaginer un cas où le motif X représente une situation (une nouvelle approche pédagogique ou un nouvel outil d'enseignement) et Y un ensemble d'erreurs que chaque élève peut commettre. Dans ce cas, la règle « si X, alors Y » peut être interprétée comme suit : lorsque la population étudiée a reçu une expérimentation (un enseignement) dans une situation modélisée par le motif X, alors on observe systématiquement le groupe d'erreurs modélisé par le motif Y. Face à ce type d'information, l'expérimentateur (l'enseignent) va être amené à repenser l'utilisation de l'approche ou l'outil modélisé par le motif X. Par contre, si le motif Y désigne un ensemble des compétences, la règle « si X, alors Y » s'interprète plutôt comme une information selon laquelle l'utilisation de l'approche ou de l'outil modélisé par X a pour conséquence l'observation des compétences modélisées par le motif Y.

Dans la pratique, il n'est pas rare de rencontrer des situations où X' n'est pas totalement inclus dans Y', c'est-à-dire qu'il existe une fraction d'objets qui échappent à la règle. Prenons par exemple, une classe de 100 élèves, supposons que 98 d'entre eux vérifient une règle « si X, alors Y », le fait d'observer 2 élèves (sur 100) qui échappent à la règle ne constitue pas une raison suffisante pour remettre en cause la règle $X \rightarrow Y$. Il est légitime de penser que le cas de ces 2 élèves est juste une exception. D'ailleurs, selon l'adage, on dit qu'il n'y a pas des règles sans exception ou que l'exception confirme la règle. La question est de savoir jusqu'où

on peut accepter cette exception. Si 2 élèves qui échappent à la règle sont considérés comme une exception, qu'en est-il pour 3 élèves ? pour 10 élèves ? etc ? Ces questions nous amènent à la notion d'implication approximative (quasi-implication) ou implication statistique.

2.4.2 Règles approximatives

Dans le cas où $X' \nsubseteq Y'$ (voir Fig. 2.1 à droite), pour analyser la validité d'une règle de type $X \to Y$, l'approche classique de R. Gras et son équipe consiste à analyser le nombre des objets qui échappent à la règle, que l'on appelle *nombre de contre-exemples.* En effet, si X' n'est pas totalement inclus dans Y', il existe parmi les objets sur lesquels on a observé les comportements modélisés par le motif X, ceux qui n'ont pas les comportements désignés par au moins une des variables constituant le motif Y (partie hachurée sur la figure 2.1 à droite). Il faut donc déterminer jusqu'où ce contre-exemple reste une exception non significative. Dans cette situation, la validité de la règle $X \to Y$, va être évaluée en fonction de la rareté des contres-exemples à la règle par rapport à la taille de la population, des cardinaux des ensembles X' et Y'. Pour mesurer mathématiquement la validité d'une règle approximative (ou quasi-règle), on a introduit le concept d'intensité d'implication, concept dont nous allons rappeler le développement dans le paragraphe ci-dessous.

Intensité d'implication Considérons le contexte $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ et considérons deux parties disjointes X et Y de $\mathcal{I}$. Pour mesurer l'intensité d'implication de la règle $X \to Y$, on commence par choisir au hasard et de manière indépendante deux parties[2] Z' et T' de $\mathcal{O}$ telles que $|Z'| = |X'|$ et $|T'| = |Y'|$ (voir Fig.2.2). On s'intéresse ensuite à la variable aléatoire

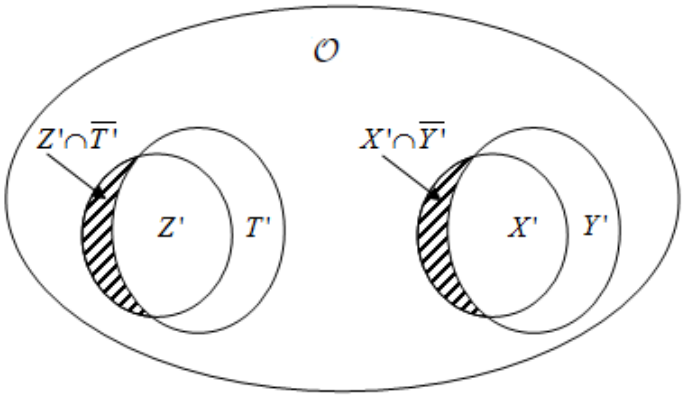


FIGURE 2.2 – Validation de la règle $X \to Y$

$Card(Z' \cap \overline{T'})$ et à l'évènement E : « $Card(Z' \cap \overline{T'}) < n_{X\overline{Y}}$ » avec $n_{X\overline{Y}} = Card(X' \cap \overline{Y'})$. Du moment que l'évènement E reste un évènement rare relativement à un seuil α très faible ($P(E) \leq \alpha$), c'est-à-dire que son évènement contraire est très fréquent ($P(\overline{E}) \geq 1 - \alpha$), le nombre de contre-exemples $n_{X\overline{Y}}$ ne peut être que relativement faible compte tenu de la composition de la population étudiée (ici et pour le reste du texte, P désigne la probabilité discrète et uniforme définie sur $(O; \mathscr{P}(O))$). La quantité $\varphi(X, Y) = 1 - P(E)$ est appelée *intensité d'implication* de la règle $X \to Y$. Si on prend une marge d'erreur très faible α

2. Pour un motif quelconque M, $M' = \{o \in \mathcal{O} / \forall\, i \in M, o(i) = 1\}$.

(souvent entre 1% et 5%), la règle $X \to Y$ est valide au niveau de confiance $1-\alpha$ lorsque $\varphi(X,Y) \geq 1-\alpha$. Remarquons que si α est nul, l'évènement E est réduit à un évènement impossible ($P(E)=0$), cela signifie que pour tout couple (Z',T') de parties de $\mathcal{O}, Card(Z' \cap \overline{T'})$ est toujours plus grand que $n_{X\overline{Y}}$. Cette situation ne peut être obtenue que lorsque $n_{X\overline{Y}}=0$ (car $Card(Z' \cap \overline{T'} \geq 0)$, c'est-à-dire lorsque $X' \subset Y'$, on retrouve donc les situations d'obtention des règles exactes.

Évaluation de l'intensité d'implication Le calcul de l'intensité d'implication dépend de la distribution de la variable aléatoire $Card(Z' \cap \overline{T'})$ qui, à son tour, dépend des hypothèses des tirages (Gras *et al.*, 2001 ; Gras et Régnier, 2009). On peut choisir une modélisation à l'issue de laquelle la variable aléatoire $Card(Z' \cap \overline{T'})$ va suivre la loi de Poisson de paramètre $\lambda = \frac{n_X n_{\overline{Y}}}{n}$ ($\lambda = n \times \frac{n_X}{n} \times \frac{n_{\overline{Y}}}{n}$), avec $n = |\mathcal{O}|$, $n_X = |X'|$ et $n_{\overline{Y}} = |\overline{Y'}|$. Dans ce cas, on a :

$$\begin{aligned} \varphi(X,Y) &= 1 - P(Card(Z' \cap \overline{T'}), \\ &= 1 - \sum_{k=0}^{n_{X\overline{Y}}} \frac{\lambda^k}{k!} e^{-\lambda}. \end{aligned}$$

Pour un $\lambda > 5$, après un centrage et une réduction, la variable aléatoire $Card(Z' \cap \overline{T'})$ va suivre une loi Gaussienne centrée et réduite $\left(\frac{Card(Z' \cap \overline{T'}) - \lambda}{\sqrt{\lambda}} \rightsquigarrow \mathcal{N}(0,1)\right)$. À partir du concept d'intensité d'implication découle un bon nombre d'outils d'analyse des données, en l'occurrence les graphes implicatifs, l'arbre hiérarchique, la notion de cohésion, etc. Toutes ces notions se trouvent à la base de la conception du logiciel CHIC (**C**lassification **H**iérarchique **I**mplicative et **C**ohésitive). Dans la pratique, il n'y a plus lieu de faire des calculs de probabilité, il suffit de connaître l'utilisation du logiciel (on peut se référer aux travaux de (Couturier et Almouloud, 2009)) et un certain minimum de bases théoriques pour l'interprétation.

2.5 Exemple d'utilisation de l'A.S.I en didactique des disciplines

Nous allons illustrer l'utilisation de l'**A**nalyse **S**tatistique **I**mplicative par le biais du logiciel CHIC pour analyser une expérimentation relative aux problématiques d'intégration pédagogique des TICE, une expérimentation effectuée dans un lycée Malagasy (Ramanantsoa et Totohasina, 2014). Avant de voir comment utiliser l'A.S.I dans ce type d'analyse, nous allons voir dans un premier temps une description plus ou moins simplifiée du contexte d'expérimentation.

2.5.1 Description du contexte d'expérimentation

Plusieurs études ont montré que l'intégration pédagogique des TIC influe positivement sur la motivation des élèves. Selon (Khvilon *et al.*, 2004, p. 88), « *Les TIC peuvent être utilisées de nombreuses façons dans les différentes branches des mathématiques, afin de motiver les élèves et de montrer l'utilité de cette discipline dans la vie quotidienne* ». Pourtant, malgré les efforts du gouvernement, des organismes non gouvernementaux (illustrés par le projet Educmad[3]) sur la promotion de l'intégration pédagogique des TICE, la principale difficulté

3. Projet de dotation des matériels informatiques dans des lycées (http ://accesmad.awdev.fr/)

reste l'insuffisance des matériels informatiques et la réticence de certains enseignants. La plupart des établissements considérés comme mieux équipés aux yeux de leurs pairs ne possèdent qu'une seule salle informatique d'une vingtaine d'ordinateurs pour l'ensemble des élèves de l'établissement et on estime par exemple que le ratio élève/ordinateur s'élève à 30. Avec un effectif moyen de 60 élèves par classe, les enseignants sont obligés de diviser la classe en plusieurs petits groupes pour pouvoir travailler dans la salle informatique. Vis-à-vis des contraintes institutionnelles relatives aux volumes de programme officiel, des programmes qui ne tiennent pas explicitement en compte l'utilisation des TIC dans la pratique enseignante, les enseignants choisissent seulement de remplir leurs obligations officielles, et par conséquent, l'intégration pédagogique des TICE reste très marginale, sinon utopique. Toujours dans le contexte éducatif malagasy, en plus des difficultés relatives à l'intégration pédagogique des TICE, on peut aussi constater le désintérêt des élèves à l'égard des disciplines scientifiques. Ces constats nous ont amené à se poser les deux questions fondamentales :

— Compte tenu des effectifs des élèves dans les établissements scolaires, comment dépasser le problème d'insuffisance des matériels informatiques pour réussir l'intégration pédagogique des TIC ?
— Comment regagner l'intérêt des élèves à l'égard des disciplines scientifiques ?

Par rapport à toutes ces problématiques, cette expérimentation a pour objectif de montrer aux enseignants réticents aux TICE (pour des raisons diverses), montrer aux décideurs qui vont choisir des stratégies pour améliorer la qualité de l'éducation que malgré ces contraintes matérielles, il est possible de faire bénéficier les élèves des plus-values que peuvent apporter les TICE (en l'occurrence la possibilité d'expérimenter, de simuler) avec un certain minimum d'équipement. Elle est fondée sur l'hypothèse selon laquelle l'utilisation « appropriée » d'un seul ordinateur et d'un vidéoprojecteur dans une salle de classe classique pourrait apporter une plus-value significative à l'égard, d'une part, de la motivation des élèves et, d'autre part, de la qualité des connaissances acquises par ces derniers. Ces plus-values vont rendre la discipline mathématique beaucoup plus attractive, beaucoup plus stimulante. Pour valider notre hypothèse, nous avons effectué une expérimentation sur l'enseignement des fonctions numériques en classe de seconde dans un lycée Malagasy.

2.5.2 Déroulement de l'expérimentation

Nous avons pris trois classes de seconde, classes de quarante élèves chacune. Nous tenons à souligner qu'avant notre expérimentation (dans cette même année ou dans les années antérieures), l'ensemble de ces élèves a été enseigné dans les mêmes conditions (même établissement, mêmes professeurs). On peut donc supposer qu'ils ont à peu près les mêmes niveaux de compétence et de connaissance avant l'expérimentation. Ensuite, nous avons utilisé deux approches pédagogiques différentes : une approche utilisant un vidéoprojecteur et un ordinateur en supplément du tableau noir (pour les deux premières classes) et une approche d'enseignement basée sur l'utilisation du tableau noir uniquement (pour la troisième classe). L'idée est d'évaluer l'ensemble de ces trois classes de la même manière (même sujet lors du contrôle) après l'enseignement des généralités sur les fonctions numériques d'une variable réelle pendant une période d'environ quatre semaines afin de pouvoir repérer l'effet de ces deux approches pédagogiques. Avant de présenter le mode d'évaluation retenu, nous allons examiner en détail les modèles d'enseignement pratiqués dans ces trois classes et procéder à la description des variables qui joueront le rôle d'indicateurs des résultats.

2.5.3 Approche pédagogique utilisée

Pour les deux premières classes, en plus de l'utilisation du tableau noir, un vidéoprojecteur et un ordinateur étaient essentiellement utilisés pour illustrer le cours, introduire une activité, animer les courbes, etc. Cette approche, ne nécessitant qu'un seul ordinateur et un vidéoprojecteur (cf. figure 2.3), a été adoptée pour répondre aux problématiques engendrées par les sureffectifs et l'insuffisance des ordinateurs dans la salle informatique, réalité observée dans plusieurs établissements scolaires de Madagascar.

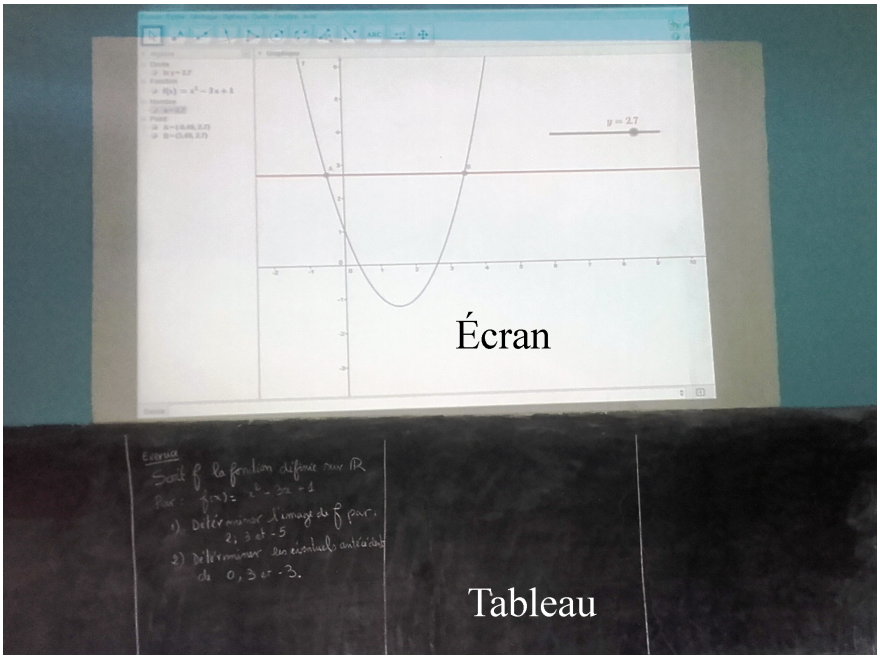


FIGURE 2.3 – Utilisation simultanée des TICE et Tableau

En effet, pour essayer d'utiliser les TIC au quotidien dans les pratiques pédagogiques, le modèle actuellement pratiqué dans le pays, illustré par le projet Educmad (cité ci-dessus) consiste à équiper la salle informatique des établissements. Ce modèle rencontre plusieurs difficultés telles que la gestion de ladite salle informatique (souvent unique dans un établissement) entre toutes les disciplines et tous les niveaux, des équipements inappropriés aux effectifs des élèves, etc. Ces difficultés sont en partie responsables de la réticence des enseignants à la pratique des TICE. C'est exactement ce constat qui nous a motivé à expérimenter une autre pratique. Par contre, pour la troisième classe, classe utilisée comme témoin, nous nous sommes contenté de recourir à la pratique pédagogique classique reposant sur l'utilisation d'un tableau noir. Soulignons que dans notre approche, c'est le caractère dynamique des objets projetés par le vidéoprojecteur et la possibilité d'utiliser ces objets dynamiques comme support d'explication qui représentent la grande différence entre les deux approches.

2.5.4 Description des variables à observer

Le choix des variables à observer dans un processus d'analyse des données dépend fortement de l'objectif de l'étude. Dans notre cas, l'objectif est d'observer et de comparer les acquis des élèves après l'expérimentation. Nous allons donc commencer ce paragraphe par

une présentation des objectifs du programme officiel (dans le système éducatif malagasy) sur l'enseignement des fonctions numériques d'une variable réelle en classe de seconde. Ces objectifs nous donneront des idées sur ce que nous devons mesurer à la fin de l'expérimentation.

Objectifs du programme officiel Initier les élèves aux démarches scientifiques afin de former l'esprit scientifique constitue l'un des objectifs de l'enseignement des mathématiques au lycée. Selon (Cariou, 2010, p. 10), « *Former l'esprit scientifique des élèves est une tâche dans laquelle se reconnaissent volontiers les enseignants scientifiques. Ils souhaitent pour cela les initier à la « démarche scientifique » ou à la « démarche expérimentale », et sont encouragés par les programmes et recommandations, aujourd'hui comme par le passé et dans de très nombreux pays* ». Dans le cas particulier de la classe de seconde du système éducatif malagasy, on peut lire dans le programme officiel que l'élève doit être capable de résoudre des problèmes conduisant à la résolution d'équations ou d'inéquations du premier et du second degré, et maîtriser la notion fondamentale de fonction numérique (image-antécédent, ensemble de définition, sens de variation, etc.). C'est pourquoi nous avons proposé d'examiner deux catégories de compétences dans cette expérimentation : compétences liées à la résolution d'un problème formel et compétences liées à la résolution d'un problème « concret ». Ces deux types de compétences seront identifiés dans deux situations que nous avons désignées respectivement par situation formelle et situation problème. Nous allons détailler les caractéristiques de ces deux types de situations.

Situation formelle Nous appelons « situation formelle », tout problème (exercice) donné sous forme de modèle mathématique sur lequel on peut tout de suite appliquer les calculs formels. En guise d'exemple, cela se produit si on demande aux élèves de résoudre l'équation $f(x) = 0$, où $f(x)$ est une expression donnée, ou encore si on demande aux élèves de résoudre graphiquement une équation ou une inéquation en partant d'une représentation graphique (cf. figure 2.4).

Exercice
On a tracé la représentation graphique d'une certaine fonction f.
Répondre aux questions ci-dessous en utilisant cette courbe.
1. Déterminer l'ensemble de définition de Df.
2. Déterminer l'image de 1 et de -2.
3. Résoudre l'équation f(x) = −2 et l'inéquation f(x) < −2.
4. Déterminer f(0).
5. Déterminer la valeur minimale de f(x).
Pour quelle valeur de x ce minimum est-il atteint ?

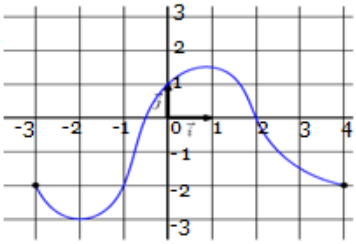


FIGURE 2.4 – Exemple d'exercice donné en « Situation formelle »

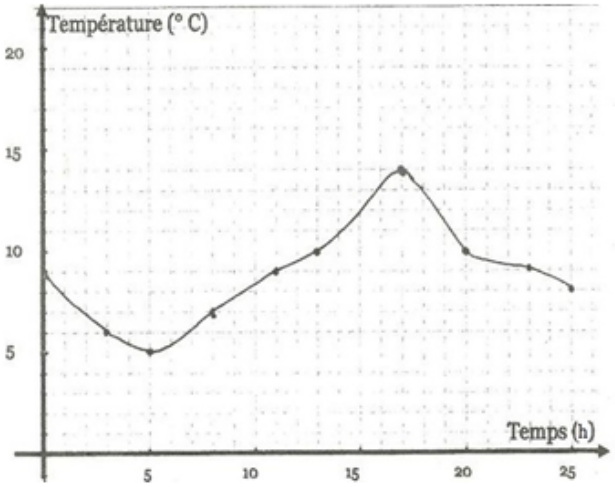
Dans la majorité des cas, ce type de problème ne demande qu'une simple application des définitions, théorèmes et propriétés connus. Au mieux, on y trouve des démonstrations de propriétés, sans trop se soucier des applications concrètes possibles.

Situation problème Par contre, nous appelons « situation problème », toute forme de problème nécessitant une modélisation mathématique avant une quelconque résolution. La modélisation fait donc partie des compétences exigées aux élèves. Pour mieux comprendre ce type de situation, reprenons les deux exemples précédemment cités (dans la situation

formelle), au lieu de demander aux élèves de résoudre une équation du type $f(x) = 0$ en leur donnant l'expression de f, on leur donne plutôt un problème conduisant à la construction de l'expression de f, et ce n'est que dans un second temps qu'on leur demande de résoudre l'équation, en leur proposant éventuellement des questions relatives au problème initialement posé. Pour le second exemple, au lieu de demander de résoudre graphiquement l'équation $f(x) = 0$ en leur donnant la représentation graphique de f, on associe plutôt la représentation graphique à un contexte différent de celui des mathématiques (social, économique, physique, etc.) et ensuite, on leur propose des problèmes se ramenant à la résolution graphique de l'équation (cf. figure 2.5).

Exercice

Un appareil a mesuré la température en un lieu, de façon continue. On a obtenu la courbe ci-après.



1- a) Donner la variation de température suivant les intervalles de temps.
b) On appelle f la fonction et t la variable temps. Dresser le tableau des variations de f.
2) Donner le maximum et le minimum de f et leurs interprétations.
3) a)A quelle(s) heure(s) la température est-elle de 10°C ?
b) Donner la valeur de la température à 3heure, à 11heure et à 17heure
c) Sur quel(s) intervalle(s) de temps la température est-elle inférieure ou égale à 7°C ?
Strictement supérieure à 10°C ? Strictement inférieur à 4°C ?

FIGURE 2.5 – Exemple d'exercice donné en « Situation problème »

Au fil des années d'enseignement, nous avons constaté que, même si les élèves arrivent à obtenir de bons résultats lorsqu'on leur propose des exercices dans la situation formelle, la plupart d'entre eux rencontrent de sérieuses difficultés face à une situation problème. Grâce aux possibilités de simulation, d'animation et d'interactivité offertes par les TICE, nous avons émis l'hypothèse selon laquelle ces outils informatiques pourraient aider les élèves à dépasser ces difficultés. Par conséquent, nous avons considéré la réussite des élèves face aux situations problèmes comme indicateur de performance de l'utilisation des TICE.

2.5.5 Mode d'évaluation et variables à observer

Après quatre semaines d'enseignement selon les deux approches précédemment citées, nous sommes passé au stade de l'évaluation en proposant les mêmes problèmes à l'ensemble des

trois classes. En effet, en évaluant l'ensemble des trois classes avec les mêmes problèmes, nous espérons identifier des caractéristiques propres à ces deux approches pédagogiques. Dans le test d'évaluation, nous avons proposé deux classes de problèmes qui nous permettent d'identifier les compétences des élèves compte tenu des deux situations que nous avons prises comme situations de référence : situation formelle et situation problème. Nous avons ensuite identifié un certain nombre d'indicateurs de compétence. À titre d'exemple, on peut se demander si l'élève est capable de chercher l'image (ou l'antécédent) d'une fonction lorsqu'on lui donne l'expression de la fonction (dans la situation formelle) et lorsqu'on lui donne un problème qui conduit à la recherche d'image et antécédent (dans la situation problème). Ces indicateurs vont jouer le rôle des variables. Ces variables vont prendre deux valeurs possibles (variables binaires) et elles vont être observées sur l'ensemble des élèves. Une variable prend la valeur 1 lorsque la compétence qu'elle identifie est présente chez l'élève, dans le cas contraire elle prend la valeur 0. Certes, l'acquisition d'une compétence peut ne pas être binaire, elle est souvent progressive. Par rapport à cela, l'utilisation des outils relatifs au contexte flous (Srikant et Agrawal, 1996 ; Totohasina, 2008) semble plus adapter à l'analyse des connaissances acquises ou non acquises. Pourtant, si on veut pousser l'analyse, on peut observer et analyser la régularité (dans le temps et dans le contexte) d'une représentation faite par un élève par rapport à une connaissance précise en décomposant cette dernière en plusieurs « atomes de connaissances » ou connaissances « élémentaires » de telle sorte que l'on puisse les observer on non chez un élève. Ces connaissances « élémentaires » vont constituer des variables binaires. Dans le cas de notre expérimentation, ces variables sont décrites dans le tableau 2.6.

Variables	**Commentaires** **Variable qui observe la réussite de l'élève dans la recherche ou la maîtrise (selon les cas) de :**
Df_Calc_For	L'ensemble de définition d une fonction en utilisant le calcul formel
Df_Sit_Pro	L'ensemble de définition d'une fonction dans une situation problème
LecG_eq_For	Solution d'équation de type $f(x) = C$ (en partant de la représentation graphique de f)
LecG_eq_Pro	Solution d'équation de type $f(x) = C$ (en partant de la représentation graphique de f donnée sous forme d'un problème)
LecG_ine_For	Solution d'inéquation de type $f(x) < C$ (en partant de la représentation graphique de f)
LecG_ine_Pro	Solution d'inéquation de type $f(x) < C$ (en partant de la représentation graphique de f donnée sous forme d'un problème)
Mait_pre_l	Solution d'équation et d'inéquation du premier degré (prérequis)
Mait_pre_2	Solution d'équation et d'inéquation du second degré (prérequis)
Util_Def_F_autr	L'écriture de la définition d'un ensemble de définition d'une fonction non rationnelle
Util_Def_F_ratio	L'écriture de la définition d'un ensemble de définition d'une fonction rationnelle
LecG_ext_For	La lecture graphique des extremums dans une situation formelle
LecG_ext_Prob	La lecture graphique des extremums dans une situation problème
LecG_ia_For	La lecture graphique des images et antécédents dans une situation formelle
LecG_ia_Prob	La lecture graphique des images et antécédents dans une situation problème
Df_LecG_For	La lecture graphique de l'ensemble de définition dans une situation formelle
Out_Tice	C'est une variable identifiant les élèves qui ont suivi l'approche d'apprentissage assistée par les TICE
Out_TN	C'est une variable identifiant les élèves qui ont suivi l'approche classique d'apprentissage (utilisation d'un tableau noir sans vidéoprojecteur)

FIGURE 2.6 – Description des variables utilisées

2.5.6 Collecte des données

À partir des productions des élèves, nous avons examiné la présence ou l'absence des informations désignées par chacune des variables décrites dans le tableau 2.6. Voici un extrait des productions d'un élève.

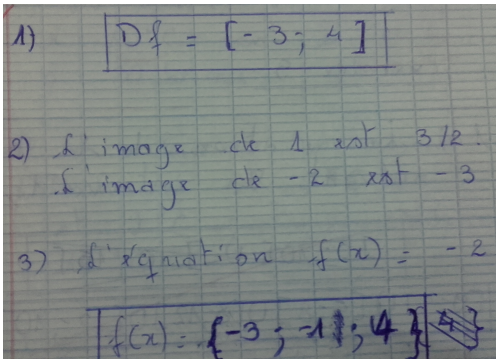


FIGURE 2.7 – Premier extrait de production

Par rapport à l'exercice présenté par la figure 2.4, en mettant de côté les problèmes de notation et de présentation, nous sommes en présence d'une copie d'un élève qui arrive à lire graphiquement :

- un ensemble de définition (Df_LecG_For=1),
- l'image d'un réel donné (LecG_ia_For=1),
- l'ensemble de solution d'une équation de type $f(x) = C$ (LecG_eq_For=1).

Notons au passage que la rédaction proposée par l'apprenant est loin d'être satisfaisante, pour que nous puissions continuer l'analyse, nous nous efforçons de comprendre ce que l'élève veut dire. Habituer ce dernier à faire une rédaction correcte constitue un autre problème dans l'enseignement des mathématiques. Chaque enseignant doit en tenir compte car rédiger en mathématiques c'est bien communiquer avec des phrases correctes (Houston, 2009) et des arguments logiques.
Un deuxième extrait de production d'un élève sur l'exercice donné en « situation problème ».

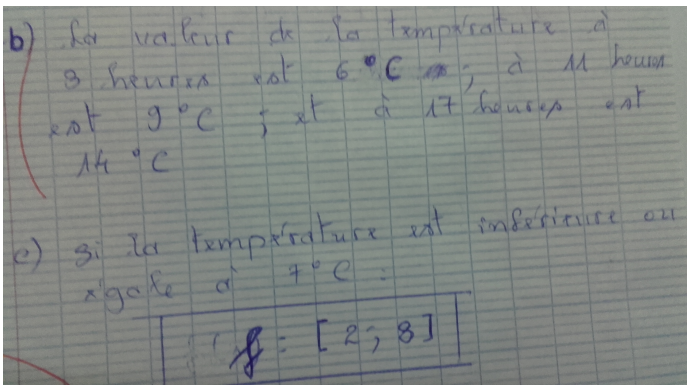


FIGURE 2.8 – Deuxième extrait de production

Dans cet extrait, on peut voir que l'élève arrive à lire graphiquement l'image d'un réel

(LecG_ia_Pro=1) et la solution d'une inéquation (LecG_eq_Pro=1).Un dernier extrait montrant la production d'un élève qui ne maîtrise pas les prérequis sur les équations du second degré (Mait_pre_2=0)et qui n'arrive pas à trouver l'ensemble de définition en manipulant les calculs formels (Df_Calc_For=0).

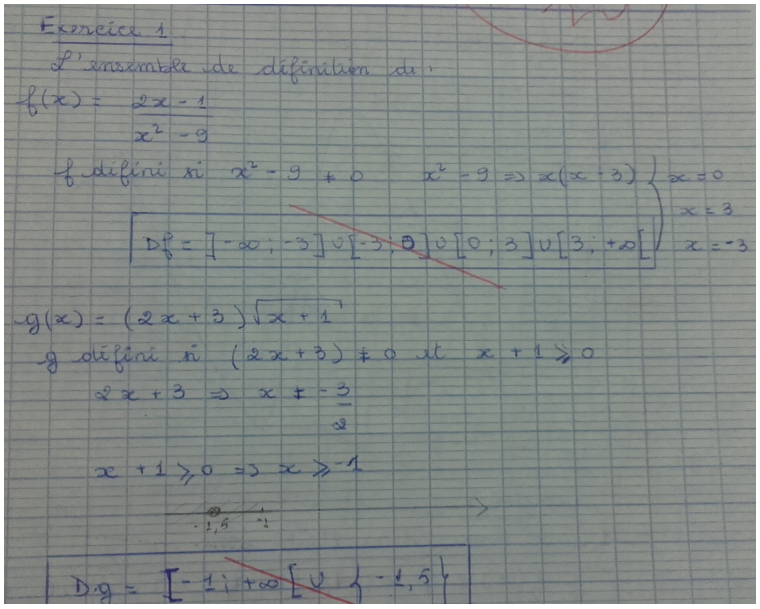


FIGURE 2.9 – Troisième extrait de production

C'est ainsi que nous avons procédé, en prenant une à une les productions des élèves et nous avons abouti au tableau binaire dont l'extrait est donné en annexe A.

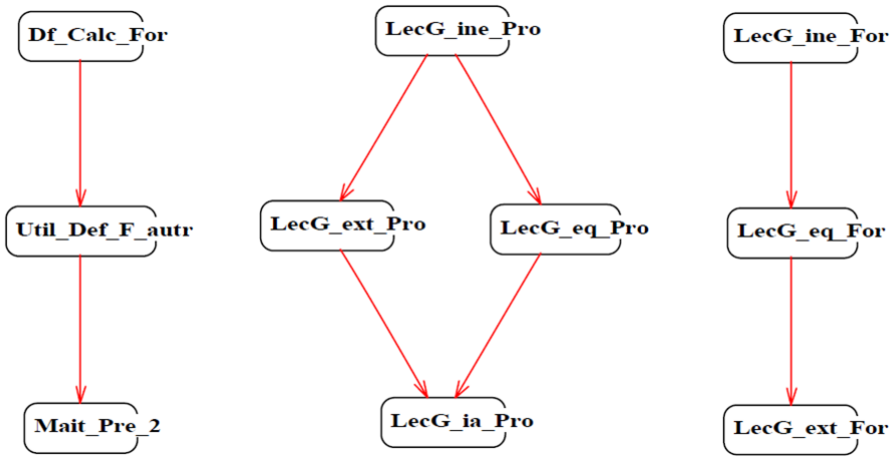
2.5.7 Outils d'analyse des données

La méthode d'analyse implicative qui vient d'être décrite dans la section 2.4 est à la base du logiciel d'analyse des données dénommé CHIC (Classification Hiérarchique Implicative et Cohésitive). Parmi les résultats fournis par CHIC, nous allons nous intéresser essentiellement au graphe implicatif et à l'arbre hiérarchique. Le graphe implicatif nous permettra d'avoir une représentation graphique de toutes les implications retenues à un seuil fixé (Gras et Régnier, 2009), tandis que l'arbre hiérarchique nous donnera des informations sur l'implication entre variables et/ou entre classes des variables (jargon utilisé en Analyse Statistique Implicative (ASI) et dans le logiciel CHIC). Dans cette analyse, nous avons utilisé quinze variables principales et deux variables supplémentaires. Les variables Out_Tice et Out_TN sont considérées comme variables supplémentaires. C'est-à-dire qu'elles ne vont pas participer à la construction du graphe implicatif, ni à celle de l'arbre hiérarchique. Elles seront utilisées pour déterminer l'approche d'enseignement qui contribue le plus à la construction d'une classe de variable. En effet, comme nos deux variables (Out_Tice et Out_TN) caractérisent le type d'enseignement reçu par les élèves, le fait de les considérer comme variables supplémentaires nous permettra de distinguer, parmi les deux types d'approche pédagogique, celle qui contribue le plus à la construction de telle ou telle classe de variables.

2.5.8 Interprétation des résultats

Après avoir repéré les quinze variables principales et deux variables supplémentaires, nous sommes passé à l'analyse des productions de ces élèves en observant la présence ou absence des connaissances ou compétences caractérisées par chacune de ces variables. Ces analyses ont abouti à un tableau de contingence croisant la liste des élèves et la liste des variables. Rappelons que les trois classes de seconde ont été évaluées avec un même sujet d'examen et l'ensemble des résultats d'analyse de production de ces élèves constitue notre base des données. Une fois que ces données sont introduites dans le logiciel CHIC, il nous donne aussitôt des résultats de traitement basé sur l'utilisation des théories sur l'analyse statistique implicative. Dans les paragraphes ci-dessous, nous allons interpréter le graphe implicatif et l'arbre hiérarchique.

Graphe implicatif En prenant un seuil $\alpha = 2\%$, c'est-à-dire une intensité d'implication (une quantité qui mesure la significativité des liens implicatifs) plus grande que 98%, nous avons obtenu le graphe implicatif ci-dessous (figure 2.10). En première lecture, nous pouvons



(a) Groupe 1 (b) Groupe 2 (c) Groupe 3

FIGURE 2.10 – Graphe implicatif (sortie du CHIC)

facilement voir que l'analyse des données de ces élèves a sorti trois réseaux des variables que nous avons appelées groupe 1, groupe 2 et groupe 3. Soulignons qu'il ne faut pas confondre les trois groupes qui viennent d'être formés et les trois classes initialement observées. Ici, c'est le logiciel CHIC (l'analyse des données) qui a réparti les variables en trois groupes. Nous pouvons aussi constater tout de suite que le groupe 2 relie exclusivement des variables relatives aux situations problèmes (définies au § 2.5.4), tandis que les deux groupes restants relient exclusivement des variables relatives aux situations formelles (définies au § 2.5.4). Maintenant, nous allons interpréter un à un les trois réseaux des variables, en prenant d'abord les deux groupes semblables (groupe 1 et groupe 3), puis nous terminerons avec le groupe 2.

Groupe 1 Selon notre lecture, ces implications s'interprètent comme suit : les élèves qui arrivent à déterminer l'ensemble de définition d'une fonction (rationnelle ou irrationnelle)

en procédant par calcul formel (Df_Calc_For) sont généralement les élèves qui savent donner la définition de l'ensemble de définition d'une fonction (Util_Def_F_autr) et les élèves qui ont une maîtrise des prérequis relatifs aux équations et inéquations de second degré. Autrement dit, ces implications montrent que la maîtrise des prérequis (équations et inéquations de second degré) et la maîtrise des définitions constituent des conditions nécessaires à la réussite de la recherche de l'ensemble de définition dans une situation formelle. Cette implication semble évidente, mais elle témoigne de l'importance de la maîtrise des définitions dans l'apprentissage des mathématiques en général.

Groupe 3 Nous sommes en présence d'un réseau de variables dont l'interprétation est plus ou moins évidente elle aussi, car les notions sont très liées (lecture graphique d'inéquation, lecture graphique d'équation et notion d'extremum). La réussite de la lecture graphique des solutions d'une inéquation est conditionnée par la réussite de la lecture graphique des équations et d'un extremum. On peut donc affirmer qu'il n'est utile de passer à la l'enseignement/apprentissage de la lecture graphique des inéquations qu'après avoir maîtrisé la lecture graphique des équations et, à son tour, la réussite de lecture graphique des images et antécédents (lecture graphique d'un extremum) conditionne la réussite de la lecture graphique des solutions d'une équation. En dehors de cet aspect évident, les deux implications montrent la hiérarchie des difficultés entre les trois notions. Dans la majorité des cas, un élève qui arrive à lire graphiquement la solution d'une inéquation arrivera à lire graphiquement la solution d'une équation et les coordonnées des extremums. Donc, lors de l'enseignement de ces notions, une attention particulière doit être accordée à la lecture graphique des solutions d'inéquations. Nous tenons à faire remarquer que la découverte de règles plus ou moins évidentes montre la cohérence et la crédibilité de notre expérimentation.

Groupe 2 Cette fois, nous sommes en présence des variables caractérisant la situation problème. En comparant les implications dans les groupes 2 et 3, on peut voir que la hiérarchie de difficulté reste la même dans les deux situations (situation formelle et situation problème). Les élèves ont donc une perception plus ou moins similaire des hiérarchies de difficulté des notions d'équation et inéquation, que ce soit dans la situation formelle ou situation problème. Par conséquent, nous avons à peu près la même interprétation que dans le groupe 3. Presque tous les élèves qui maîtrisent la résolution graphique des inéquations réussissent la lecture graphique des solutions d'équations, des coordonnées d'extremums et, évidemment, la lecture graphique de l'image et de l'antécédent. On peut aussi interpréter les règles par contraposées ; rappelons que logiquement, pour toutes variables binaires a et b, (a implique b) est équivalente à (non b implique non a). Donc, selon notre graphe implicatif, la non-maîtrise de la lecture graphique des images et antécédents pourrait entraîner la non-maîtrise du reste. En général, un enseignant doit tenir compte de ces hiérarchies de difficulté, d'une part dans l'évaluation (en n'évaluant plus les items de niveau de difficulté un peu plus bas, après avoir évalué un item se trouvant à un niveau de difficulté un peu plus haut), d'autre part dans le processus d'enseignement proprement dit (en faisant attention au passage d'un item à l'autre, le fait de sauter un item non maîtrisé pouvant entraîner un blocage chez les élèves pour le reste du processus).

Arbre Hiérarchique Nous pouvons remarquer que dans l'interprétation du graphe implicatif, nous n'avons pas utilisé les deux variables (Out_TN, Out_Tice) qui sont directement

liées à notre principale préoccupation. Remarquons d'abord que ces deux variables caractérisent le type d'approche pédagogique utilisée dans l'expérimentation : la variable Out_TN spécifie l'approche utilisant seulement le tableau noir ; par contre la variable Out_Tice caractérise l'approche utilisant le vidéoprojecteur en supplément du tableau noir. Comme nous l'avons précisé au troisième paragraphe, nous avons statué ces deux variables comme étant des variables « supplémentaires ». Ces variables seront utilisées dans l'interprétation de l'arbre hiérarchique pour déterminer leurs contributions dans la formation des classes. Soulignons qu'avec le logiciel CHIC, on peut savoir le pourcentage de la contribution d'une variable supplémentaire à la construction d'une classe. Il faut noter que l'arbre hiérarchique confirme et complète l'analyse du graphe implicatif (Ottaviani et Zannoni, 2001). Il la complète dans le sens où il donne des implications entre variables et règles, ou encore entre plusieurs règles (figure 2.11). Mais, conformément à notre principale préoccupation, dans cette interprétation, nous allons plutôt nous intéresser à l'interprétation des variables supplémentaires. Avant

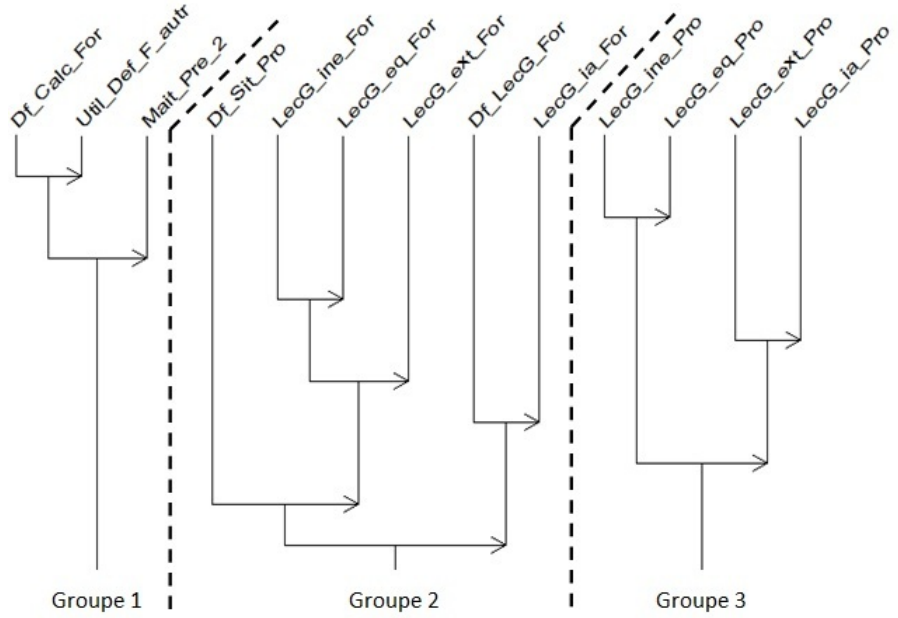


FIGURE 2.11 – Arbre hiérarchique implicatif et cohésitif

tout, constatons la formation des classes compatibles avec les trois groupes que nous venons d'identifier dans l'interprétation de graphe implicatif. Cette représentation graphique est donc compatible avec les résultats que nous avons avancés dans l'interprétation du graphe implicatif. Analysons maintenant la contribution de nos deux variables supplémentaires. Pour cela, nous avons reproduit ci-dessous les résultats d'analyse de contribution relatifs à nos trois groupes, résultat fourni par CHIC. La figure 2.12 représente une copie d'écran de CHIC, figure que nous allons interpréter. Premièrement, remarquons la très forte contribution de la variable Out_TN, variable caractérisant l'approche pédagogique sans utilisation d'outils informatiques, sur la construction du groupe 2 (groupe caractérisé par les variables relatives à l'évaluation des compétences face aux problèmes donnés dans des situations formelles). Cette variable contribue à la construction de ce groupe (de cette classe selon le jargon employé dans CHIC) avec un risque très faible. C'est à dire, d'après ce résultat, les élèves qui ont eu une approche d'enseignement non assisté par les TICE contribuent le plus dans la construc-

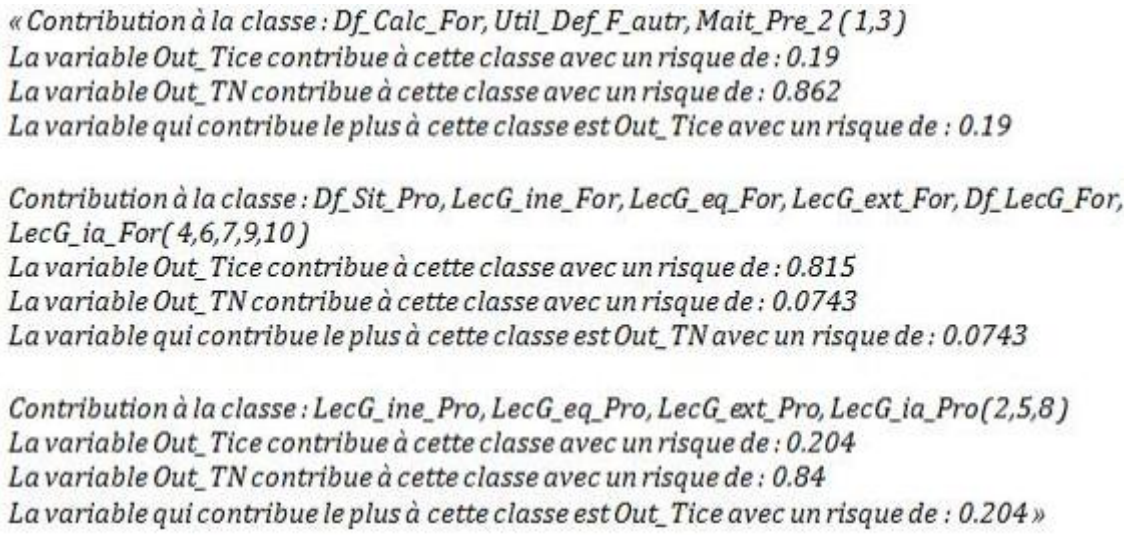


FIGURE 2.12 – Contributions des variables supplémentaires

tion des classes de variables identifiées par le groupe 2, variables utilisées pour évaluer des compétences sur la résolution des problèmes formels (situation formelle). Par contre, c'est plutôt la variable Out_Tice, variable caractérisant l'approche pédagogique avec utilisation d'outils informatiques qui contribue le plus à la construction du groupe 3, classe de variable relative à l'évaluation des compétences sur les exercices donnés en situation problème. Sur l'échelle de un, le risque 0,20 peut paraître grand, mais si on regarde de plus près les deux implications qui constituent le groupe, on peut s'apercevoir que la contribution de Out_Tice dans la construction de ces deux règles est respectivement de 0,19 et de 0,008. On peut donc affirmer qu'à l'issue de l'expérimentation, c'est l'approche pédagogique caractérisée par la variable Out_Tice qui est responsable de la construction de cette classe (groupe 3). En ce qui concerne le groupe 1, groupe formé par des variables relatives au calcul formel, on a encore une fois une forte contribution de la variable Out_Tice. Compte tenu des deux premières tendances, on peut affirmer que l'approche pédagogique utilisant les outils informatiques (un ordinateur et un vidéoprojecteur) contribue fortement à la réussite des élèves face aux exercices donnés sous forme de problèmes (situation problème). Par contre, l'approche n'utilisant que le tableau noir contribue plutôt à la réussite des élèves face au calcul formel (situation formelle). En observant la formation du groupe 1, on constate que l'approche utilisant les outils informatiques peut aussi contribuer à la réussite des élèves face aux problèmes formels. Cette expérimentation montre donc que la combinaison « équilibrée » de ces deux approches a amélioré la qualité d'enseignement.

2.6 Conclusion de l'expérimentation

Cette expérimentation nous a montré un exemple concret de l'utilisation de l'Analyse Statistique Implicative (ASI) en didactique de mathématiques. Nous avons pu mettre des liens entre des approches pédagogiques, les outils utilisés et les connaissances acquises par les élèves. L'étude des contributions des variables supplémentaires a justifié l'hypothèse selon laquelle l'exploitation des caractères dynamiques des animations vidéoprojetées (caractérisés par la variable Out_Tice) est la principale responsable de la construction des classes des variables montrant des liens des compétences face aux « situations problèmes ». De son côté, l'approche pédagogique exploitant seulement le tableau noir (caractérisée par la variable Out_TN) constitue la principale raison de la construction des liens entre les variables observées en situation formelle. On peut tirer au moins deux résultats de cette expérimentation. D'une part, il est possible de mettre en évidence les liens de cooccurence qui peuvent s'interpréter comme des liens de cause à effet entre les erreurs et les compétence des élèves. Ces informations seront utiles pour orienter les travaux de remédiation après un ou plusieurs échecs d'un élève. D'autre part, nous avons pu conclure aussi qu'il est conseillé d'utiliser simultanément le vidéoprojecteur et le tableau noir si l'on veut que nos élèves puissent pousser leurs raisonnements grâce à la manipulation des objets dynamiques et en même temps, si l'on souhaite développer chez eux la faculté de manipuler formellement les objets mathématiques. Entre les catégories d'enseignants qui veulent céder totalement la place du tableau noir à une projection vidéo et ceux qui s'accrochent à l'ancienne pratique et qui refusent toutes réformes en matière de technologies éducatives, il faut savoir trouver le juste équilibre. La vidéoprojection est efficace pour les observations et les conjectures, mais un concept mathématique ne s'apprend qu'en manipulant les objets mathématiques.

2.7 Autres mesures utilisées dans l'extraction des règles

La fiabilité des règles extraites dépend fortement de la qualité de la mesure utilisée. C'est une raison pour laquelle un certain nombre des chercheurs dans le domaine d'Extraction des Connaissances à partir des Données (ECD) ont concentré leurs efforts sur l'étude des qualités des mesures. On peut citer entre autres les travaux de (Blanchard *et al.*, 2004) sur l'intensité d'implication entropique, une extension de l'intensité d'implication proposée par l'équipe de R. Gras afin de quantifier les déséquilibres entre l'exemple et le contre-exemple d'une règle et à la fois de sa contraposée. Les travaux de (Lenca *et al.*, 2003) sur les critères d'évaluation des mesures de qualité des règles d'association et les travaux de (Le Bras *et al.*, 2010) sur la robustesse des règles extraites. Depuis, plusieurs mesures ont été proposées dans la littérature. Une liste de 61 mesures est dressée dans (Grissa, 2013) lors d'une étude comportementale des mesures de qualité. Certaines d'entre elles sont données dans le tableau 2.2. Pour que les expressions des mesures puissent avoir des sens, dans le tableau 2.2 et pour le reste de ce rapport, toutes les mesures ne sont effectuées que sur des règles de type $X \rightarrow Y$ avec $|X'|$ et $|Y'|$ non nuls, X' et Y' représentent respectivement les extensions de X et de Y.

Noms	Expressions
Support	$P\left(Y' \cap X'\right)$
Confiance	$P\left(Y'/X'\right)$
Lift	$\frac{P(Y'/X')}{P(Y')}$
M_{GK}	$\frac{P(Y'/X')-P(Y')}{1-P(Y')}$, $si\ P\left(Y'/X'\right) > P\left(Y'\right)$ $\frac{P(Y'/X')-P(Y')}{P(Y')}$, $si\ P\left(Y'/X'\right) \leq P\left(Y'\right)$
Rappel	$P\left(X'/Y'\right)$
Conviction	$\frac{P(X')P(\overline{Y'})}{P(X' \cap \overline{Y'})}$
Dépendance	$\left\lvert P\left(\overline{Y'}\right) - P\left(\overline{Y'}/X'\right)\right\rvert$

Tableau 2.2 – Quelques mesures de qualité d'une règle $X \rightarrow Y$

Selon (Vaillant, 2006 ; Vaillant *et al.*, 2005), les mesures de qualités peuvent classer très différemment les règles et la sélection des *bonnes connaissances* passe par l'utilisation d'une *bonne* mesure. Nous allons voir dans le paragraphe suivant, la limite de l'utilisation du couple de mesure le plus utilisé dans les algorithmes d'extraction des règles, à savoir « Support-Confiance ».

2.8 Limite de l'utilisation du couple support-confiance

Parmi les mesures disponibles dans la littérature, notamment celles qui sont citées dans le tableau 2.2, le couple support-confiance se trouve à la base de plusieurs algorithmes d'extraction des règles d'association. Une règle $X \rightarrow Y$ est valide selon le couple support-confiance lorsque son support et sa confiance dépassent respectivement les valeurs minimales $minSupp$ et $minConf$ fixées par l'utilisateur. Comme en témoignent les nombreuses études sur les propriétés et la qualité des mesures, entre autres celles de (Vaillant *et al.*, 2004 ; Le Bras *et al.*, 2010), la fiabilité d'une règle extraite dépend de la qualité de la mesure utilisée. Malgré l'utilisation plus ou moins généralisée du couple support-confiance, ces deux mesures ont suscité beaucoup de critiques de la part des chercheurs dans le domaine d'ECD (Vaillant, 2006 ; Lallich et Teytaud, 2004). En effet, ces deux mesures peuvent prendre des valeurs élevées même dans une situation d'indépendance et de répulsion. Pour illustrer ce propos, nous allons considérer des contextes fictifs rapportés dans les tableaux de contingence ci-après (cf. tableau 2.3).
Faisons maintenant quelques mesures relatives aux règles $X \rightarrow Y$ et $Z \rightarrow T$. Après calcul, nous avons obtenu les valeurs ci-dessous :
En prenant un $minSup$ et un $minConf$ égal à $0,7$ et en utilisant le couple Support-Confiance, la règle $(X \rightarrow Y)$ du premier contexte de le tableau 2.3 (à gauche) est bien une règle valide. Pourtant, si on examine de plus près les deux motifs X et Y, on peut facilement se rendre compte qu'ils sont stochastiquement indépendants $(P(Y'/X') = P(Y'))$. Donc,

(a) Indépendance

	X	$\overline{X}$	$\sum$
Y	560	240	800
$\overline{Y}$	140	60	200
$\sum$	700	300	1000

(b) Répulsion

	Z	$\overline{Z}$	$\sum$
T	600	200	800
$\overline{T}$	180	20	200
$\sum$	780	220	1000

Tableau 2.3 – Exemples de situations indésirables validées par Support-Confiance

(a) **Pour la règle** $(X \rightarrow Y)$

$$\begin{aligned} \text{Conf}(X \rightarrow Y) &= 0{,}8 \\ \text{Supp}(Y) &= 0{,}8 \\ \text{Supp}(X) &= 0{,}7 \\ P(Y'/X') &= P(Y') \end{aligned}$$

(b) **Pour la règle** $(Z \rightarrow T)$

$$\begin{aligned} \text{Conf}(Z \rightarrow T) &= 0{,}76 \\ \text{Supp}(Z) &= 0{,}78 \\ \text{Supp}(T) &= 0{,}80 \\ P(T'/Z') &< P(T') \end{aligned}$$

l'affirmation selon laquelle la présence du motif X influence celle du motif Y est erronée, pourtant selon le couple Support-Confiance, elle est acceptable avec un niveau de confiance 70% ($minConf = 0,7$). C'est pareil pour le deuxième contexte décrit dans le tableau 2.3 (à droite), Conf$(Z \rightarrow T)$ et Supp$(Z \rightarrow T)$ dépassent respectivement le $minConf$ et le $minSupp$. Donc selon ce couple de mesures, la règle $(Z \rightarrow T)$ est bien une règle valide. Autrement dit, Support et Confiance ont validé délibérément la règle $Z \rightarrow T$ alors que les deux motifs Z et T se défavorisent mutuellement $(P(T'/Z') < P(T'))$, c'est-à-dire qu'il est logiquement plus intéressant d'étudier plutôt la règle $Z \rightarrow \overline{T}$ que la règle $Z \rightarrow T$.
Plus important encore, ces constats influencent la définition même d'une règle redondante. Rappelons que selon (Pasquier, 2000b ; Bastide *et al.*, 2002), entre deux règles de même Confiance, de même Support et qui ont des conséquents comparables, celle qui a un conséquent plus grand (au sens de l'inclusion) est plus informative ; c'est-à-dire, entre $r_1 : X \rightarrow Y$ et $r_2 : X \rightarrow Z$, avec $Z \subset Y$, la règle r_1 est plus informative et, la règle r_2 est considérée comme une règle redondante. En effet, pour tous motifs X, Y, Z d'un contexte quelconque $\mathcal{K}$ tels que $Z \subset Y$, on a toujours :

$$\begin{aligned} n_{XZ} &\geq n_{XY}\,(Z \cap X \subseteq Y \cap X)\,, \\ \text{Conf}(X \rightarrow Z) &\geq \text{Conf}(X \rightarrow Y), \\ \text{Donc, } \text{Conf}(X \rightarrow Y) \geq minConf &\Rightarrow \text{Conf}(X \rightarrow Z) \geq minConf. \end{aligned}$$

Par rapport à cette logique, si $(X \rightarrow Y)$ est valide, on pourrait croire que pour tout $Z \subset Y$, $(X \rightarrow Z)$ est toujours valide. Nous allons voir, via la figure 2.13 qu'il existe des situations dans lesquelles ce raisonnement n'est pas valide. Considérons un contexte binaire d'extraction $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$, prenons trois motifs X, Y et Z tels que $Z \subset Y$. D'après la propriété d'antimonotonie de support, on a : Supp$(Z) \geq$ Supp(Y). Représentons maintenant dans la figure 2.13 les extensions des motifs X, Y et Z $(|X'| = n_X,\ |Y'| = n_Y,\ |Z'| = n_Z\ et\ |\mathcal{O}| = n)$

Dans cette figure, nous avons fait en sorte que la Confiance de $X \rightarrow Y$ soit égale à celle de $X \rightarrow Z$ et que le Support de Z soit supérieur à celui de Y. En supposant que la règle $X \rightarrow Y$ soit valide au sens de la mesure confiance (Figure 2.13 à gauche), la règle $X \rightarrow Z$ ne le sera

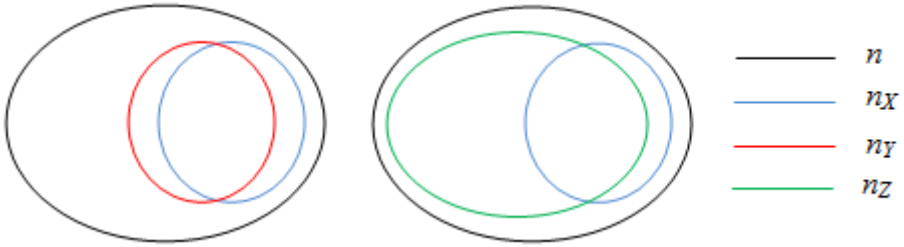


FIGURE 2.13 – Validité de $X \to Z$ et $X \to Y$, avec $Y \subset Z$

pas. En effet, si on observe bien la figure 2.13 (à droite), on peut constater que la validité de la règle $\overline{X} \to Z$ est de loin la plus plausible que celle de la règle $X \to Z$ (le co-occurrence de $\overline{X}Z$ est largement plus grand que celui de XZ) . On peut donc affirmer que même si la règle $(X \to Y)$ est valide, il peut y avoir un motif Z dans $\mathscr{P}(\mathcal{I})$ tel que $(X \to Z)$ ne soit pas valide. Déduire la validité d'une règle $r_2 : X \to Z$ à partir de celle de $r_1 : X \to Y$ en se basant seulement sur le fait que Z soit une partie de Y et du fait que la Confiance de $(X \to Z)$ soit plus grande ou égale à celle de $(X \to Y)$ peut conduire à une information (connaissance) erronée.
Le nombre trop important des règles qui pourraient être extraites constitue une autre limite de l'utilisation du couple support-confiance. En effet, si on fixe les seuils $minSupp$ et $minConf$ assez petits, ces deux mesures vont valider des règles ayant le Support et la Confiance qui dépassent leurs seuils respectifs même dans des situations d'indépendance ou de répulsion. Par conséquent, dans la plupart du temps, le nombre des règles extraites est trop élevé. Les utilisateurs vont être noyés dans ce nombre exorbitant de règles dont certaines sont inintéressantes (dans le cas d'indépendance et de répulsion). Les règles intéressantes seront cachées par les règles non intéressantes. D'un autre côté, si on fixe un $minSupp$ et un $minConf$ trop élevés, non seulement on peut perdre des règles ayant un petit support et qui peuvent avoir une Confiance élevée (les pépites de connaissances), mais aussi, rien ne garantit la suppression des règles non intéressantes. Enfin, le couple Support-Confiance peut valider des règles dans la situation de répulsion (La prémisse défavorise le conséquent). C'est-à-dire qu'à elles seules, ces deux mesures ne permettent pas de découvrir les règles négatives. Ceci explique l'afflux des règles positives extraites par ce couple de mesures. Donc, pour garantir la fiabilité des règles, par la suite, la fiabilité des connaissances extraites d'une base des données, il est nécessaire d'utiliser d'autres mesures de qualité. Soulignons que la mesure Support (Support d'un motif) reste très utilisée parce qu'elle permet de réduire considérablement l'espace de recherche.
Dans notre étude, nous avons privilégié l'utilisation du couple de mesures Support-M_{GK}. Nous allons voir dans le prochain paragraphe les raisons qui nous ont poussé à faire ce choix.

2.9 Choix de la mesure M_{GK}

L'utilité des règles extraites dépend de la qualité des mesures utilisées. Les recherches sur l'amélioration des qualités des règles extraites ont abouti aux propositions de plusieurs mesures de qualité. Nous venons de voir la limite de l'utilisation du couple Support et Confiance, pourtant, Support et Confiance figurent parmi les mesures les plus utilisées dans les algo-

rithmes d'extraction des règles d'association. Cependant, sur les 61 mesures énumérées dans (Grissa, 2013), si l'utilisateur envisage d'utiliser des mesures autres que le Support et la Confiance, le choix reste problématique, en particulier pour les non spécialistes du domaine. On peut voir dans (Guillaume *et al.*, 2010) une liste de 21 propriétés souhaitables d'une mesure de qualité. On peut aussi montrer qu'aucune de ces mesures ne vérifient pas l'ensemble des propriétés souhaitables et les utilisateurs se retrouvent dans une sorte d'embarras de choix face à la question « laquelle ou lesquelles utiliser » ? Dans notre cas, nous avons choisi d'utiliser le couple des mesures Support et M_{GK}. Afin de donner les raisons qui nous ont poussé à faire ce choix, rappelons la définition d'une mesure normalisée donnée dans (Totohasina *et al.*, 2004) et (Diatta *et al.*, 2007).

Définition 2.4 (Mesures normalisées)**.** *Soit $X \to Y$ une règle d'association. Une mesure de qualité μ est dite normalisée si elle vérifie les cinq conditions ci-dessous :*

1. $\mu(X \to Y) = -1$ *si et seulement si X et Y sont incompatibles ;*
2. $-1 < \mu(X \to Y) < 0$ *si et seulement si X défavorise Y, ou de manière équivalente, X et Y sont négativement dépendants ;*
3. $\mu(X \to Y) = 0$ *si et seulement si X et Y sont indépendants ;*
4. $0 < \mu(X \to Y) < 1$ *si et seulement si X favorise Y, ou de manière équivalente, X et Y sont positivement dépendants ;*
5. $\mu(X \to Y) = 1$ *si et seulement si X implique logiquement Y ;*

Vérifions maintenant que M_{GK} est une mesure normalisée selon la définition 2.4. Par définition, pour motif X, Y d'un contexte binaire d'extraction $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$:

$$M_{GK}(X \to Y) = \begin{cases} \frac{P(Y'/X') - P(Y')}{1 - P(Y')}, & \text{si } X \text{ favorise } Y, \\ \frac{P(Y'/X') - P(Y')}{P(Y')}, & \text{si } X \text{ défavorise } Y. \end{cases} \tag{2.1}$$

Notation Dans la relation (2.1), la composante favorisante est souvent notée par M^f_{GK} et la composante défavorisante par M^d_{GK}, c'est-à-dire, pour deux motifs X et Y, on a :

$$\begin{aligned} M_{GK}(X \to Y) = M^f_{GK}(X \to Y) &= \tfrac{P(Y'/X') - P(Y')}{1 - P(Y')}, \ \textit{si } X \text{ favorise } Y, \\ M_{GK}(X \to Y) = M^d_{GK}(X \to Y) &= \tfrac{P(Y'/X') - P(Y')}{P(Y')}, \ \textit{si } X \text{ défavorise } Y. \end{aligned}$$

Dans le reste du rapport, nous utiliserons la notation $\mathrm{M}_{GK}(X \to Y)$ pour désigner $M^f_{GK}(X \to Y)$ ou $M^d_{GK}(X \to Y)$ selon la nature de lien entre les motifs X et Y. Pour tous motifs X, Y de $\mathscr{P}(\mathcal{I})$, $P(Y'/X') \leq 1$, donc $P(Y'/X') - P(Y') \leq 1 - P(Y')$ et, par conséquent :

$$\frac{P(Y'/X') - P(Y')}{1 - P(Y')} \leq 1.$$

Dans le cas où X favorise Y (i. e. $P(Y'/X') > P(Y')$), on a : $\dfrac{P(Y'/X') - P(Y')}{1 - P(Y')} > 0$. En conclusion, si X favorise Y, on a :

$$0 < M_{GK}(X \to Y) \leq 1. \tag{2.2}$$

D'autres part, pour tous motifs X, Y de $\mathscr{P}(\mathcal{I})$, $P(Y'/X') \geq 0$, donc $P(Y'/X') - P(Y') \geq -P(Y')$. En divisant les deux membres de cette inégalité par $P(Y')$, on obtient :

$$\frac{P(Y'/X') - P(Y')}{P(Y')} \geq -1.$$

Dans le cas de X défavorise Y(i. e. $P(Y'/X') \leq P(Y')$), on a :

$$\frac{P(Y'/X') - P(Y')}{P(Y')} \leq 0.$$

C'est-à-dire, dans le cas de répulsion mutuelle, on a :

$$-1 \leq M_{GK}(X \to Y) \leq 0. \tag{2.3}$$

Dans le cas d'incompatibilité entre deux motifs X, Y de $\mathscr{P}(\mathcal{I})$ $(X' \cap Y' = \emptyset)$, on ne peut être que dans le cas de répulsion mutuelle et comme $P(Y'/X') = 0$ on a :

$$M_{GK}(X \to Y) = -1. \tag{2.4}$$

Dans le cas d'indépendance entre la prémisse et le conséquent, $(P(Y'/X') = P(Y'))$, on a :

$$M_{GK}(X \to Y) = \frac{P(Y'/X') - P(Y')}{P(Y')} = 0. \tag{2.5}$$

Enfin, dans le cas d'implication logique $(X' \subset Y')$, on a toujours $P(Y'/X') = 1$ donc

$$M_{GK}(X \to Y) = \frac{P(Y'/X') - P(Y')}{1 - P(Y')} = 1 \tag{2.6}$$

Les relations (2.3) à (2.6) montre que la mesure M_{GK} prend des valeurs particulières dans les situations de références autres que la situation d'équilibre. Ces propriétés prouvent que cette mesure est bien normalisée selon la définition 2.4. La figure 2.14 résume les valeurs particulières de M_{GK} dans les situations de référence.

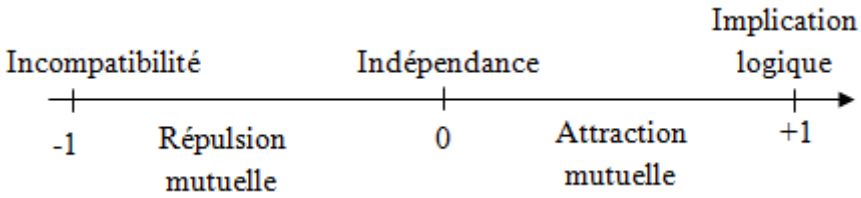


FIGURE 2.14 – Situations de références

Sur les 21 propriétés énumérées par Guillaume *et al.* en 2010, la mesure M_{GK} vérifie 13. De plus, les propriétés mathématiques de cette mesure sont compatibles avec la propriété sur l'équivalence entre le statut d'une implication logique et de sa contraposée. En effet, étant données deux propositions P et Q, selon les propriétés classiques en logique formelle, on n'a pas d'équivalence entre les propositions $(P \Rightarrow Q)$ et $(\overline{P} \Rightarrow \overline{Q})$. Par contre, on en a une entre $(P \Rightarrow Q)$ et $(\overline{Q} \Rightarrow \overline{P})$. Ce constat nous amène à affirmer que d'une part, contrairement

à la 18e propriété décrite dans (Guillaume *et al.*, 2010), une mesure doit être capable de différencier les deux règles $(P \Rightarrow Q)$ et $(\overline{P} \Rightarrow \overline{Q})$ parce que, logiquement ces deux règles ne sont pas équivalentes. D'autre part, si on veut qu'une mesure garde la propriété d'équivalence entre une proposition et sa contraposée, elle doit attribuer systématique la même valeur pour les deux règles de types $(P \Rightarrow Q)$ et $(\overline{Q} \Rightarrow \overline{P})$. Encore une fois, nous sommes en présence d'un point fort de la mesure M_{GK}. En effet, selon la propriété 3.6 (cf. page 41), pour tous motifs P et Q, on a toujours :

$$\begin{cases} M_{GK}\left(P \to Q\right) & = & M_{GK}\left(\overline{Q} \to \overline{P}\right), \text{ Si } P \text{ favorise } Q, \\ M_{GK}\left(P \to \overline{Q}\right) & = & M_{GK}\left(Q \to \overline{P}\right), \text{ Si } P \text{ défavorise } Q. \end{cases}$$

Selon la propriété et le corollaire 4.1(cf. page 75), on a toujours une équivalence entre la validité selon M_{GK} d'une règle et de sa contraposée (que ce soit dans le cas d'attraction mutuelle ou de répulsion). Les règles valides selon la mesure M_{GK} vérifient la propriété de logique classique sur l'équivalence d'une proposition et de sa contraposée. Ensuite, la relation entre la mesure M_{GK} et le test d'indépendance de χ^2, établie par Totohasina *et al.* en 2004 permet de trouver objectivement les valeurs critiques de M_{GK}, valeurs de références pour valider ou non une règle au niveau de confiance fixé par l'utilisateur. Ces valeurs critiques permettent aussi de garantir (selon l'approche de test d'indépendance de χ^2) la fiabilité de dépendance stochastique entre les variables intervenant dans une règle valide. C'est-à-dire qu'une règle valide selon M_{GK} ne peut être qu'une règle entre deux motifs stochastiquement dépendants. Autrement dit, la mesure M_{GK} ne validera jamais une règle entre deux motifs indépendants selon le test d'indépendance de χ^2 relativement à un niveau de confiance fixé par l'utilisateur. De plus, par le fait qu'on a deux expressions différentes de M_{GK} selon la nature de dépendance, attraction ou répulsion, entre prémisse et conséquent, on n'a aucun risque de mélanger la nature de dépendance. Dans le cas de répulsion entre deux motifs X et Y, on n'a qu'à s'intéresser aux motifs X et $\overline{Y}$. Enfin, le concept de normalisation repris dans (Totohasina, 2008) permet d'unifier un certain nombre de mesures. En effet, dans ces travaux, on a déjà énuméré 19 mesures M_{GK}-normalisables. Une mesure est dite M_{GK}-normalisable s'il existe une application qui transforme les valeurs de cette mesure en valeurs de M_{GK}. Cette mesure joue donc un rôle d'unificateur de mesures. Ces propriétés mathématiques et le concept de normalisation constituent quelques raisons qui nous ont poussé à choisir cette mesure. En ce qui concerne la mesure Support, on l'a gardée pour une raison purement algorithmique. En effet, son utilisation permet de restreindre le champ de recherche sur les motifs qui ont une certaine fréquence de présence dans les objets (les transactions). Cette fréquence est tout à fait modulable selon le choix de l'utilisateur en variant le seuil *minSupp*. La mesure Support a l'avantage d'être facile à interpréter. Elle représente la fréquence de la présence des motifs dans la transaction. Il est donc facile pour les utilisateurs de fixer le *minSupp* selon leurs besoins.
Dans le but d'avoir une valeur fixe dans la situation d'équilibre, c'est-à-dire dans la situation où le nombre de contre-exemples est égal au nombre d'exemples $(n_{X\overline{Y}} = n_{XY})$, et d'exclure certaines règles apriori non inintéressantes, en 2010, Guillaume *et al.* ont défini une nouvelle version de la mesure M_{GK} (cf. définition 2.5).

Définition 2.5. *Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte binaire d'extraction. Pour tous motifs X, Y*

de $\mathscr{P}(\mathcal{I})$, les nouvelles versions de M_{GK} *sont données par les expressions ci-après.*

$$M_{G_1}(X \to Y) = \begin{cases} \frac{P(Y'/X') - \max\left(P(Y'), \frac{1}{2}\right)}{1 - \max\left(P(Y'), \frac{1}{2}\right)}, & si\ X\ favorise\ Y, \\ \frac{P(Y'/X') - \max\left(P(Y'), \frac{1}{2}\right)}{\max\left(P(Y'), \frac{1}{2}\right)}, & si\ X\ défavorise\ Y. \end{cases}$$

$$M_{G_2} = \begin{cases} \frac{P(Y'/X') - \min\left(P(Y'), \frac{1}{2}\right)}{\min\left(P(Y'), \frac{1}{2}\right)}, & si\ P\left(Y'/X'\right) < \min\left(P\left(Y'\right), \frac{1}{2}\right), \\ 0, & si\ \min\left(P\left(Y'\right), \frac{1}{2}\right) \leq P\left(Y'/X'\right) \leq \max\left(P\left(Y'\right), \frac{1}{2}\right), \\ \frac{P(Y'/X') - \max\left(P(Y'), \frac{1}{2}\right)}{1 - \max\left(P(Y'), \frac{1}{2}\right)}, & si\ \max\left(P\left(Y'\right), \frac{1}{2}\right) < P\left(Y'/X'\right). \end{cases}$$

Remarquons que pour les motifs de supports plus grands que $0,5$, les valeurs données par les deux mesures M_{G_1} et M_{GK} coïncident. De plus, avec la version originelle, grâce à sa relation avec le test d'indépendance de χ^2 et à l'utilisation des valeurs critiques dans la validation des règles, on ne risque pas de valider une règle dans des situations indésirables, c'est-à-dire en cas d'équilibres ou de dépendance non significative entre les motifs).
Par ailleurs, selon la proposition 2.1 ci-dessous, avec cette nouvelle version, la mesure M_{GK} perd la qualité d'être implicatif, une qualité tant souhaitée en ASI.

Proposition 2.1. *Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte binaire d'extraction. Il existe X, Y de $\mathscr{P}(\mathcal{I})$ tels que $M_{G_1}(X \to Y) \neq M_{G_1}(\overline{Y} \to \overline{X})$. C'est-à-dire que la composante favorisante de la mesure M_{G_1} n'est plus implicative.*

Preuve.
En effet, si X favorise Y (i. e. $P(Y'/X') > P(Y')$), alors dans le cas où $P(Y') < \frac{1}{2}$, on a :

$$M_{G_1}(X \to Y) = \frac{P\left(Y'/X'\right) - \frac{1}{2}}{1 - \frac{1}{2}} = 2P\left(Y'/X'\right) - 1.$$

Dans ce même cas, $M_{G_1}\left(\overline{Y} \to \overline{X}\right) = \frac{P\left(\overline{X}'/\overline{Y}'\right) - \max\left(P\left(\overline{X}'\right), \frac{1}{2}\right)}{1 - \max\left(P\left(\overline{X}'\right), \frac{1}{2}\right)}$. Cette dernière expression varie selon la valeur de $P(\overline{X}')$ et elle n'a rien à voir avec $2P\left(Y'/X'\right) - 1$.
Donc, en général, si X favorise Y, on n'a aucune raison pour égaliser $M_{G_1}(X \to Y)$ et $M_{G_1}\left(\overline{Y} \to \overline{X}\right)$. D'où la composante favorisante de M_{G_1} n'est pas implicative. □

Ainsi, le choix entre la nouvelle et la version originelle de M_{GK} s'impose. Donc, nous avons travaillé avec la version originelle de M_{GK} à composante favorisante implicative associée à la mesure Support.

2.10 Conclusion partielle

Nous avons vu qu'il existe plusieurs mesures de qualité pour valider un lien implicatif entre deux motifs et que l'on peut utiliser ces outils mathématiques pour valider les liens de cause à effet dans la recherche en didactique de mathématiques. Ces mesures ont été définies de différentes manières, elles ont par conséquent des propriétés mathématiques différentes. L'intensité d'implication par exemple est définie à partir de la comparaison des nombres de contre-exemples à la règle étudiée. Face à ce nombre trop importante des mesures de qualité,

il n'est pas évident pour un utilisateur non spécialiste du domaine de choisir laquelle ou lesquelles utiliser. Malgré les critiques à l'encontre des mesures Support et Confiance, ces deux mesures figurent parmi les plus utilisées dans les algorithmes d'extraction des règles d'association. Grâce à ses propriétés mathématiques et à son rôle dans la normalisation des autres mesures de qualités, nous avons choisi d'utiliser la mesure M_{GK} dans sa version originelle. Après les étapes d'extraction des règles, la phase d'interprétation est incontournable pour valider une quelconque connaissance utile et exploitable. Si on est en présence d'un nombre trop important des règles, la phase d'interprétation peut devenir très vite problématique dans le sens où l'utilisateur sera obligé de fouiller dans l'ensemble des règles valides, celles qui sont plus informatives. Il est donc nécessaire de restreindre l'ensemble des règles présentées à l'utilisateur pour interprétation. Évidemment, il faut faire en sorte que cette restriction ne fasse perdre aucune information utile. D'où la notion des bases des règles d'association que nous allons développer dans les prochains chapitres.

Chapitre 3

Extraction des bases des règles d'association

Sommaire

3.1 Introduction

Étant donné le nombre qui pourrait être très élevé des règles valides possibles, il est facile d'imaginer la difficulté, voire l'impossibilité d'exploitation de ces dernières. Ce problème a préoccupé l'esprit des nombreux chercheurs dans le domaine des fouilles des données et, plusieurs techniques et solutions ont été proposées au fil des années. À titre d'exemple, on peut citer les approches orientées utilisateurs (Klemettinen *et al.*, 1994) comme l'utilisation des templates, une pratique consistant à proposer aux utilisateurs de spécifier dans des templates les items qui vont apparaître dans les prémisses et les conséquents, ou encore, en créant une taxonomie sur les items. Ces types d'approches peuvent réduire considérablement le nombre de règles valides sauf que d'une part, elles ne permettent pas de faire découvrir les règles inattendues (puisque les items ont été choisis ou ont subi des contraintes) et d'autre part, elles ne garantissent pas la réduction des règles redondantes (règles qui n'apportent aucune information supplémentaire). Comme en témoigne un bon nombre des travaux dans la littérature (Pasquier, 2000a ; Bastide *et al.*, 2002 ; Le Floc'h *et al.*, 2003 ; Gasmi *et al.*, 2006 ; Hamrouni *et al.*, 2005, 2011), l'extraction des bases des règles est une solution à ce problème de surabondance des règles extraites. Par définition, une base des règles d'association est un sous-ensemble des règles non redondantes à partir duquel, par l'utilisation des axiomes d'inférence, il est possible d'extraire l'ensemble de toutes les règles valides. La définition d'une base des règles dépend des axiomes d'inférence permettant de retrouver les autres règles et ces derniers sont liés étroitement à l'expression des mesures de qualité utilisées. Nous avons pu constater qu'une bonne partie des algorithmes d'extraction des bases des règles sont conçus

avec le couple de mesures support et confiance. Dans ce chapitre, nous allons faire un tour d'horizon des principaux travaux qui vont nous servir à décrire et à comprendre l'extraction des bases des règles d'association. Nous commencerons par donner quelques définitions et propriétés qui vont servir à la description des bases. Ensuite, nous allons voir séparément les bases des règles valides selon le couple de mesures Support-Confiance et celles qui sont valides selon le couple Support-M_{GK}.

3.2 Quelques définitions et propriétés sur les fermetures

Nous allons voir dans cette section les quelques définitions et propriétés qui vont nous servir pour définir et comprendre les concepts utilisés dans l'extraction des règles d'association. Nous soulignons que la totalité des propriétés données dans cette section sont des propriétés connues, nous les avons rappelé et prouvé dans le but d'une part, de mieux présenter l'état de l'art sur l'extraction des bases des règles d'association, en particulier les bases des règles M_{GK}-valides et, d'autre part, pour faciliter la compréhension de leurs utilisations dans les présentations de nos contributions qui feront l'objet du prochain chapitre.

Définition 3.1 (Opérateur de fermeture)**.** *Soit $(E, \leq)$ un ensemble ordonné. Une application γ de $(E, \leq)$ dans $(E, \leq)$ est un opérateur de fermeture si elle possède les trois propriétés suivantes :*

1. *Isotonie : $\forall X, Y \in E,\ X \leq Y \Rightarrow \gamma(X) \leq \gamma(Y)$;*
2. *Extensivité : $\forall X \in E,\ X \leq \gamma(X)$;*
3. *Idempotence : $\forall X \in E,\ \gamma(\gamma(X)) = \gamma(X)$.*

Définition 3.2 (Correspondance de Galois)**.** *Soient $(E, \leq)$, $(F, \leq)$ deux ensembles ordonnés, $f : E \to F$ et $g : F \to E$ deux applications. Le couple (f, g) sera dit correspondance de Galois entre E et F si, pour tous $e_1, e_2 \in E$ et pour tous $y_1, y_2 \in F$, les trois conditions suivantes sont vérifiées :*

1. *Antitonie : $e_1 \leq e_2$ implique $f(e_2) \leq f(e_1)$*
2. *Antitonie : $y_1 \leq y_2$ implique $g(y_2) \leq g(y_1)$*
3. *Extensivité : $e_1 \leq g \circ f(e_1)$ et $y_1 \leq f \circ g(y_1)$.*

Propriété 3.1. *Considérons le contexte binaire d'extraction $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$. Soit ϕ une application qui associe une partie X de $\mathcal{I}$ à son extension X' de $\mathcal{O}$.*

$$\begin{array}{rcl} \phi : \mathscr{P}(\mathcal{I}) & \longrightarrow & \mathscr{P}(\mathcal{O}) \\ X & \longmapsto & X' = \phi(X) = \{o \in \mathcal{O} / \forall x \in X, (o, x) \in \mathcal{R}\} \end{array}$$

Soit ψ une application qui associe une partie X' de $\mathcal{O}$ à son intension X de $\mathcal{I}$.

$$\begin{array}{rcl} \psi : \mathscr{P}(\mathcal{O}) & \longrightarrow & \mathscr{P}(\mathcal{I}) \\ X' & \longmapsto & X = \psi(X') = \{i \in \mathcal{I} / \forall o \in X', (o, x) \in \mathcal{R}\} \end{array}$$

Le couple d'applications (ϕ, ψ) définit une correspondance de Galois entre l'ensemble des parties de $\mathcal{O}$ et l'ensemble des parties de $\mathcal{I}$.

Preuve. $(\mathscr{P}(\mathcal{I}), \subseteq)$ et $(\mathscr{P}(\mathcal{O}), \subseteq)$ sont des ensembles ordonnés.
Soit $X_1, X_2 \in \mathscr{P}(\mathcal{I})$ tels que $X_1 \subseteq X_2$. D'après la définition de ϕ, on a :
$\phi(X_1) = \{o \in \mathcal{O} / \forall x \in X_1, (o,x) \in \mathcal{R}\}$ et $\phi(X_2) = \{o \in \mathcal{O} / \forall x \in X_2, (o,x) \in \mathcal{R}\}$.
Soit o un élément de $\phi(X_2)$, pour tout $x \in X_2$, $(o,x) \in \mathcal{R}$. Or, tous les éléments de X_1 sont des éléments de X_2 ($X_1 \subseteq X_2$). Donc, pour tout x dans X_1, $(o,x) \in \mathcal{R}$. Autrement dit, o est un élément de $\phi(X_1)$. $\phi(X_2)$ est donc une partie de $\phi(X_1)$, d'où l'antitonie de ϕ. Le même raisonnement appliqué à ψ pourra montrer l'antitonie de ψ.
Prenons maintenant un élément X de $\mathscr{P}(\mathcal{I})$. Soit x un élément de X, selon la définition de ϕ, pour tout o dans $\phi(X)$, $(o,x) \in \mathcal{R}$. Selon la définition de ψ, x est un élément de $\psi(\phi(X))$; donc, $X \subseteq \psi(\phi(X))$. L'application $\psi \circ \phi$ est donc extensive. Le même raisonnement peut être appliqué à $\phi \circ \psi$. On peut donc conclure que le couple (ϕ, ψ) définit une correspondance de Galois. □

Propriété 3.2. *Dans un contexte binaire $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$, désignons par (ϕ, ψ) une correspondance de Galois entre $\mathscr{P}(\mathcal{O})$ et $\mathscr{P}(\mathcal{I})$. Les application $\gamma = \psi \circ \phi$ et $\gamma' = \phi \circ \psi$ sont des opérateurs de fermeture définis respectivement sur $\mathscr{P}(\mathcal{I})$ et $\mathscr{P}(\mathcal{O})$.*

Preuve.
Isotonie
Soient X et Y deux motifs tels que $X \subset Y$. En se servant de l'antitonie de ϕ et ψ, on obtient :

$$\begin{aligned} X \subset Y &\Rightarrow \phi(X) \supset \phi(Y), \\ \phi(X) \supset \phi(Y) &\Rightarrow \psi(\phi(X)) \subset \psi(\phi(Y)), \\ \text{donc, } X \subset Y &\Rightarrow \gamma(X) \subset \gamma(Y). \end{aligned}$$

Extensivité
Soit x_k un élément de X, montrons que x_k est aussi un élément de $\gamma(X)$.

$$\begin{aligned} \gamma(X) &= \psi \circ \phi(X) \\ &= \psi(\phi(X)) \\ &= \{i \in \mathcal{I} / \forall o \in \phi(X), (o,i) \in \mathcal{R}\}. \end{aligned}$$

Donc $\gamma(X)$ est formé par les items en relation avec tous les objets de $\phi(X)$. Or, par définition, $\phi(X) = \{o \in \mathcal{O} / \forall x \in X, (o,x) \in \mathcal{R}\}$. $\phi(X)$ est formé par des objets dont chacun est en relation avec tous les éléments de X. C'est à dire qu'un x quelconque de X est en relation avec tous les objets de $\phi(X)$. En prenant un x_k dans X, on a : $x_k \in \mathcal{I}, \forall o \in \phi(X), (o, x_k) \in \mathcal{R}$, x_k est donc un élément de $\gamma(X)$ et on conclut que X est une partie de $\gamma(X)$.
Idempotence
Soit X un motif et posons $\begin{cases} O = \phi(X) \\ Z = \psi(O) \end{cases}$ et cherchons $\gamma(X) = \psi \circ \phi(X)$:

$$\begin{aligned} \gamma(X) &= \psi \circ \phi(X) \\ &= \psi(O) \\ &= Z. \end{aligned}$$

$\phi(X)$ représente l'ensemble des objets en relation avec tous les items composants le motif X ($O = \phi(X) = \{o \in \mathcal{O} / \forall x \in X, (o,x) \in \mathcal{R}\}$). Autrement dit, pour chaque objet o de O, o est

en relation avec tous les composants du motif X.
Par définition, $\psi(O) = \{i \in \mathcal{I} / \forall o \in O, (o,i) \in \mathcal{R}\}$. Z est donc composé des éléments en relation avec tous les objets de O. C'est à dire, pour chaque item z de Z, z est en relation avec tous les composants de O. Cherchons maintenant $\gamma(\gamma(X))$.

$$\begin{aligned} \gamma(\gamma(X)) &= \gamma(Z) \\ &= \psi \circ \phi(Z) \end{aligned}$$

Par définition, $\phi(Z) = \{o \in \mathcal{O} / \forall z \in Z, (o,z) \in \mathcal{R}\}$. Montrons que $\phi(Z)$ et O sont égaux.
Soit o_k un élément de $\phi(Z)$, c'est-à-dire : $\forall z \in Z, (o_k, z) \in \mathcal{R}$. Comme $X \subseteq \gamma(X)$, (i. e. $X \subseteq Z$), une propriété vraie pour tout $z \in Z$ est toujours vraie pour tout $x \in X$ (X n'est qu'une partie de Z). Donc, pour tout $x \in X, (o_k, x) \in \mathcal{R}$. o_k est donc un élément de $\phi(X)$ qui n'est autre que O. On aboutit à : $\phi(Z) \subseteq O$. Prenons maintenant un objet o_k de O, o_k est en relation avec tous les éléments de Z ($\psi(O) = Z$). o_k est un objet de $\mathcal{O}$ en relation avec tous les composants de Z, o_k est donc un élément de $\psi(Z)$. On en déduit que $O \subseteq \phi(Z)$. Tout compte fait, on a prouvé que O et $\phi(Z)$ sont égaux.
Donc, $\gamma(\gamma(X)) = \gamma(Z) = \psi \circ \phi(Z) = \psi(O) = Z = \gamma(X)$. D'où la propriété d'idempotence. □

Propriété 3.3. *Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte binaire d'extraction, O_1, O_2 deux parties de $\mathcal{O}$ et I_1, I_2 deux parties de $\mathcal{I}$. Supposons que (ψ, ϕ) soit une correspondance de Galois sur $\mathcal{K}$, nous avons les propriétés suivantes :*

$$\begin{aligned} I_1 \subseteq I_2 &\Rightarrow \phi(I_2) \subseteq \phi(I_1), \\ O_1 \subseteq O_2 &\Rightarrow \psi(O_2) \subseteq \psi(O_1), \\ \phi \circ \psi(\phi(X)) &= \phi(X), \\ O_1 \subseteq \phi(I_1) &\Leftrightarrow I_1 \subseteq \psi(O_1). \end{aligned}$$

Preuve.
Soit I_1, I_2 deux motifs de $\mathscr{P}(\mathcal{I})$ tels que $I_1 \subseteq I_2$. Prenons un o_2 dans $\phi(I_2)$, montrons que $o_2 \in \phi(I_1)$. Par définition $\phi(I_2) = \{o \in \mathcal{O} / \forall x \in I_2, (o,x) \in \mathcal{R}\}$ et $o_2 \in \phi(I_2)$ signifie que pour tout x dans I_2, $(o_2, x) \in \mathcal{R}$. Or tous les éléments de I_1 sont des éléments de I_2, cela signifie qu'une propriété vraie pour tous les éléments de I_2 reste vraie pour tous les éléments de I_1 (puisque $I_2 \supseteq I_1$). Par conséquent, à partir d'un o_2 dans $\phi(I_2)$, on peut constater que : $\forall x \in I_1,\ (o_2, x) \in \mathcal{R}$. Cela signifie que $o_2 \in \phi(I_1)$, ce qui prouve que $\phi(I_2) \subseteq \phi(I_1)$.
Appliquons le même raisonnement, mais cette fois avec ψ. Partant d'un i_2 dans $\psi(O_2)$, montrons que i_2 est aussi dans $\psi(O_1)$. $i_2 \in \psi(O_2)$ signifie que pour tout o dans O_2, $(o, i_2) \in \mathcal{R}$. Or tous les éléments de O_1 sont des éléments de O_2 (puisque $O_1 \subseteq O_2$), donc pour tout o dans O_1, $(o, i_2) \in \mathcal{R}$. Cela signifie que i_2 est un élément de $\psi(O_1)$, donc $\psi(O_2) \subseteq \psi(O_1)$.
Pour tout motif X de $\mathscr{P}(\mathcal{I})$, $X \subseteq \gamma(X)$ et en utilisant l'antitonie de ϕ on obtient :
$\phi(X) \supseteq \phi(\gamma(X))$, c'est-à-dire :

$$\phi(X) \supseteq \phi(\psi \circ \phi(X)). \tag{3.1}$$

En utilisant l'extensivité de γ' définie par : $\gamma' = \phi \circ \psi$, pour tout motif X de $\mathscr{P}(\mathcal{I})$ ($\phi(X) \in \mathscr{P}(\mathcal{O})$), on a : $\phi(X) \subseteq \gamma'(\phi(X))$, c'est-à-dire :

$$\phi(X) \subseteq \phi \circ \psi(\phi(X)). \tag{3.2}$$

En combinant (3.1) et (3.2), on obtient l'égalité ci-après :
pour tout X dans $\mathscr{P}(\mathcal{I}), \phi \circ \psi(\phi(X)) = \phi(X)$. Enfin, supposons que $O_1 \subseteq \phi(I_1)$, cela signifie que pour tout o dans O_1, $o \in \phi(I_1)$. Or, $o \in \phi(I_1)$ est équivalent à : pour tout i dans I_1, $(o, i) \in \mathcal{R}$.
En prenant donc un item i dans I_1 et sous l'hypothèse $O_1 \subseteq \phi(I_1)$, on a toujours $(o, i) \in \mathcal{R}$, et cela pour tout $o \in O_1$. Donc, selon la définition de $\psi(O_1)$, l'item i appartient toujours à $\psi(O_1)$. Donc, pour tout O_1 dans $\mathscr{P}(\mathcal{O})$ et pour tout I_1 dans $\mathscr{P}(\mathcal{I})$: $O_1 \subseteq \phi(I_1) \Rightarrow I_1 \subseteq \psi(O_1)$.
Effectuons le même raisonnement pour montrer l'implication réciproque. Prenons un objet quelconque o d'une partie O_1 de $\mathscr{P}(\mathcal{O})$, montrons que o est toujours dans $\phi(I_1)$ du moment qu'on a l'hypothèse $I_1 \subseteq \psi(O_1)$.
Cette hypothèse signifie que pour tout i dans I_1, i est toujours dans $\psi(O_1)$. Or, i dans $\psi(O_1)$ signifie que pour tout o dans O_1, $(o, i) \in \mathcal{R}$. Maintenant, prenons un élément quelconque o dans O_1. Selon l'hypothèse, pour tout $i \in I_1$, $(o, i) \in \mathcal{R}$. Ceci prouve que o est élément de $\phi(I_1)$. Donc on peut affirmer que $O_1 \subseteq \phi(I_1)$. D'où l'implication réciproque. □

Définition 3.3 (Motif fermé)**.** *Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte binaire d'extraction. Une partie X de $\mathcal{I}$ est appelée motif fermé, si sa fermeture $\gamma(X)$ est égale à elle-même ($\gamma(X) = X$).*

Propriété 3.4. *Soit X un motif fermé et notons par $[X]$ l'ensemble des motifs ayant la même fermeture que X : $[X] = \{Y \in \mathscr{P}(\mathcal{I}) / \gamma(Y) = \gamma(X)\}$ où γ représente l'opérateur de fermeture défini dans le contexte $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$.*

1. *X est l'unique fermé dans $[X]$*
2. *$\forall Y \in \mathscr{P}(\mathcal{I}), Y \in [X] \Rightarrow Y \subset X$*
3. *$\forall Y \in \mathscr{P}(\mathcal{I}), Supp(Y) = Supp(\gamma(Y))$*
4. *$\forall Y \in \mathscr{P}(\mathcal{I}), Y \in [X] \Rightarrow Supp(Y) = Supp(X)$*

Preuve.
Soit Y un élément de $[X]$ et supposons que Y soit un fermé quelconque de $\mathscr{P}(\mathcal{I})$:

$$\begin{cases} \gamma(Y) = \gamma(X) = X, \\ \gamma(Y) = Y. \end{cases} \tag{3.3}$$

Le système (3.3) nous conduit nécessairement à $Y = X$. Pour le deuxième point, si on prend un élément quelconque Y de $\mathscr{P}(\mathcal{I})$, on a : $Y \subset \gamma(Y)$ (γ est extensive). Comme $Y \in [X]$, on a $\gamma(Y) = \gamma(X) = X$, donc $Y \subset X$. Pour tout motif I, on a :

$$\begin{aligned} \text{Supp}(\gamma(I)) &= \frac{\text{Card}(\phi(\gamma(I)))}{\text{Card}(\mathcal{O})} \\ &= \frac{\text{Card}(\phi(\psi \circ \phi(I)))}{\text{Card}(\mathcal{O})} \\ &= \frac{\text{Card}(\phi(I))}{\text{Card}(\mathcal{O})} \text{(Propriété 3.3)} \\ &= \text{Supp}(I). \end{aligned}$$

Le quatrième point est une conséquence immédiate du troisième. □

Propriété 3.5. *Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte binaire d'extraction et γ un opérateur de fermeture. Pour tous motifs X, Y de $\mathscr{P}(\mathcal{I})$, on a toujours :*

$$\gamma(X \cup Y) = \gamma(\gamma(X) \cup \gamma(Y)).$$

Preuve.
Prenons deux motifs X, Y de $\mathscr{P}(\mathcal{I})$. En utilisant la propriété d'extensivité de l'opérateur de fermeture, on a :

$$\begin{cases} X \subseteq \gamma(X) \\ Y \subseteq \gamma(Y) \end{cases} \Rightarrow X \cup Y \subseteq \gamma(X) \cup \gamma(Y).$$

À partir de cette implication et en utilisant l'isotonie de γ, on obtient :

$$\gamma(X \cup Y) \subseteq \gamma(\gamma(X) \cup \gamma(Y)). \tag{3.4}$$

D'autre part, pour tous motifs X, Y, on a toujours $X \subseteq X \cup Y$ et $Y \subseteq X \cup Y$. Ensuite, l'isotonie γ nous donne les inclusions : $\gamma(X) \subseteq \gamma(X \cup Y)$ et $\gamma(Y) \subseteq \gamma(X \cup Y)$ et par la suite : $\gamma(X) \cup \gamma(Y) \subseteq \gamma(X \cup Y)$. En utilisant de nouveau l'isotonie et l'idempotence de γ, on obtient :

$$\gamma(\gamma(X) \cup \gamma(Y)) \subseteq \gamma(X \cup Y). \tag{3.5}$$

En combinant 3.4 et 3.5, on obtient l'égalité :

$$\gamma(X \cup Y) = \gamma(\gamma(X) \cup \gamma(Y)).$$

□

Définition 3.4 (Générateur minimal)**.**
Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte binaire d'extraction et X un motif fermé. On appelle générateur de X, le plus petit (au sens de l'inclusion) motif G_X dont la fermeture est égale à X.

Définition 3.5 (Pseudo-fermé)**.**
Désignons par γ l'opérateur de fermeture associé au contexte binaire $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$. Une partie X de $\mathcal{I}$ est appelée pseudo-fermée ou γ-critique, si elle contient la fermeture de tous ces sous-ensembles qui sont des pseudo-fermés.

Exemple 3. *Le tableau 3.1 donne quelques catégories de motifs du contexte binaire $\mathcal{K}$ du tableau 2.1.*

Fermés	Pseudo-fermés	Générateurs
$\emptyset, A, B, D$		
AB, AD, BE, BD		E, C, F
ABE, ABC, ABD, BDE	$E, C, F, ABDE$	AE, DE, CE
$ABEF, ABCE, ABCD$		CF, DF, CD
$ABCEF, ABDEF, ABCDEF$		

Tableau 3.1 – Différentes catégories de motifs

Propriété 3.6 (Quelques propriétés de M_{GK})**.**
Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte binaire d'extraction, X et Y deux parties de $\mathcal{I}$.

P1 *Si X défavorise Y, alors X favorise $\overline{Y}$ et réciproquement.*

P2 *Si X favorise Y, alors $\mathrm{M}_{GK}(X \to Y) = \mathrm{M}_{GK}(\overline{Y} \to \overline{X})$: M_{GK} est favorablement implicative.*

P3 *Si X défavorise Y (X favorise $\overline{Y}$), alors $\mathrm{M}_{GK}(X \to \overline{Y}) = \mathrm{M}_{GK}(Y \to \overline{X})$.*

Preuve.
Supposons que X défavorise Y

$$\begin{array}{rcl} P(Y/X) & < & P(Y) \\ 1 - P(Y/X) & > & 1 - P(Y) \\ P(\overline{Y}/X) & > & P(\overline{Y}) \end{array}$$

On a donc X favorise $\overline{Y}$. De la même manière, en supposant que X favorise $\overline{Y}$, c'est-à-dire $P(\overline{Y}/X) > P(\overline{Y})$, on aboutit facilement à $P(Y/X) < P(Y)$. Supposons maintenant que X favorise Y, en se servant de **P1**, on peut déduire que $\overline{Y}$ favorise $\overline{X}$. Exprimons maintenant $\mathrm{M}_{GK}(\overline{Y} \to \overline{X})$ en fonction de $\mathrm{M}_{GK}(X \to Y)$.

$$\begin{array}{rcl} \mathrm{M}^f_{GK}(\overline{Y} \to \overline{X}) & = & \dfrac{P(\overline{X'}/\overline{Y'}) - P(\overline{X'})}{1 - P(\overline{X'})} \\ & = & \dfrac{1 - P(X'/\overline{Y'}) - P(\overline{X'})}{P(X')} \\ & = & \dfrac{P(X') - \frac{P(X')}{P(\overline{Y'})} P(\overline{Y'}/X')}{P(X')} \\ & = & \dfrac{1 - P(Y') - P(\overline{Y'}/X')}{1 - P(Y')} (\text{en simplifiant avec } P(X')) \\ & = & \dfrac{P(Y'/X') - P(Y')}{1 - P(Y')} \\ & = & \mathrm{M}^f_{GK}(X \to Y) \end{array}$$

Enfin, supposons que X défavorise Y.

Selon [**P1**], X défavorise Y est équivalent à X favorise $\overline{Y}$, donc :

$$\begin{aligned}
M^f_{GK}\left(X \to \overline{Y}\right) &= \frac{P\left(\overline{Y'}/X'\right) - P\left(\overline{Y'}\right)}{1 - P\left(\overline{Y'}\right)} \\
&= \frac{1 - P\left(Y'/X'\right) - P\left(\overline{Y'}\right)}{P\left(Y'\right)} \\
&= \frac{P\left(Y'\right) - P\left(Y'/X'\right)}{P\left(Y'\right)} \\
&= \frac{P\left(Y'\right) - \frac{P(Y')P(X'/Y')}{P(X')}}{P\left(Y'\right)} \\
&= 1 - \frac{P\left(X'/Y'\right)}{P\left(X'\right)} \\
&= \frac{P\left(X'\right) - P\left(X'/Y'\right)}{P\left(X'\right)} \\
&= \frac{1 - P\left(\overline{X'}\right) - \left(1 - P\left(\overline{X'}/Y'\right)\right)}{\left(1 - P\left(\overline{X'}\right)\right)} \\
&= \frac{P\left(\overline{X'}/Y'\right) - P\left(\overline{X'}\right)}{1 - P\left(\overline{X'}\right)} \\
&= M^f_{GK}\left(Y \to \overline{X}\right).
\end{aligned}$$

□

Une conséquence immédiate de la propriété **P1** fait que les études sur les règles positives et négatives peuvent être effectuées en se servant de la seule composante favorisante de M_{GK}, noté M^f_{GK}. Dans la suite de notre texte, nous avons essentiellement utilisé cette composante favorisante, que ce soit dans la mesure des règles positives ou négatives. Quant à la propriété **P2**, elle nous montre qu'on peut toujours déduire (si l'on veut) la mesure des règles négatives bilatérales à partir des règles positives correspondantes.

3.3 Bases des règles

Le problème d'extraction des règles d'association à partir des données peut être divisé en deux grandes parties : l'extraction des motifs fréquents à partir des données et l'extraction des règles à partir des motifs fréquents. À son tour, l'extraction des règles d'association à partir des motifs fréquents a rencontré deux gros problèmes : la complexité des algorithmes d'extraction des motifs fréquents (Salleb, 2003) et le nombre prohibitif des règles extraites dont une grande majorité sont redondantes [1]. En effet, pour une base de données de dimension [2] m, le nombre des motifs fréquents possibles est égal à $2^m - 1$ et, pour un motif fréquent Y de taille k, le nombre des règles possibles (de type $X \to Y \backslash X$ avec $X \subset Y$) est égal à $2^k - 2$. Pour illustrer ce problème de redondance, considérons un contexte binaire d'extraction $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ au sein duquel le motif $Y = ABCD$ est fréquent (le nombre des objets ou

1. Apportent les mêmes informations ou des informations moins pertinentes que d'autres règles.
2. Ici la dimension désigne le nombre des attributs.

transactions contenant Y dépasse un minimum ($minSupp$) fixé par l'utilisateur). Soulignons que les sous-ensembles d'un motif fréquent sont forcément des motifs fréquents. Examinons le type d'information apportée par les règles ci-dessous.

$$\begin{array}{lrcl} r_1: & A & \rightarrow & BCD \\ r_2: & AB & \rightarrow & CD \\ r_3: & AC & \rightarrow & BD \\ r_4: & AD & \rightarrow & BC \\ r_5: & ABC & \rightarrow & D \\ r_6: & ABD & \rightarrow & C \end{array}$$

À un seuil fixé par l'utilisateur, supposons que ces six règles soient valides. Comparons les informations fournies par la première et la sixième règles.

$r_1: A \rightarrow BCD$ Dans ce contexte binaire d'extraction, si le motif A est présent dans une transaction (ou dans un objet), alors on observe souvent, sinon toujours la présence du motif BCD dans cette même transaction.

$r_6: ABC \rightarrow D$ Si le motif ABC fait partie d'une transaction, on observe souvent ou toujours la présence du motif D dans cette même transaction.

Selon la première règle, il suffit que A soit présent dans une transaction pour affirmer avec une forte probabilité la présence du motif BCD. Par contre, selon la sixième règle, il faut la présence de tous les items A, B et C (composants du motif ABC) avant de pouvoir déduire la présence du motif D. Il est donc clair (sous l'hypothèse de la validité de ces deux règles) que la sixième règle n'apporte rien de nouveau par rapport à la première règle. C'est ce type de règle que l'on appelle « Règle redondante ». Dans le présent exemple, les cinq dernières règles n'apportent aucune information supplémentaire par rapport à l'information délivrée par la première règle. Si on extrait toutes les règles valides relativement à un seuil de validité quelconque, on risque de tomber sur un nombre trop élevé des règles dont la plupart pourraient être redondantes. Pour se convaincre du nombre très élevé des règles possible, nous allons dénombrer ces dernières pour une base des données de dimension m.

Nombre de règles possibles

Considérons un contexte binaire $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ avec : $\mathcal{O} = \{o_1, o_2, ..., o_n\}$ ensemble des n objets, $\mathcal{I} = \{i_1, i_2, ..., i_m\}$ ensemble des m attributs. Nous allons dénombrer le nombre des règles possibles de type $X \rightarrow Y \backslash X$ avec $X, Y \in \mathscr{P}(\mathcal{I})$ et $X \subset Y$.
D'abord, ici Y ne peut pas être un singleton.

Pour Y dans la partie de $\mathcal{I}$ à 2 éléments

Pour les C_m^2 Y possibles, on a :

$$\begin{cases} Y = \{i_1,\, i_2\} \to X \in \mathscr{P}\left(\{i_1,\, i_2\}\right) \;:\; C_2^1 \text{ règles possibles,} \\ Y = \{i_1,\, i_3\} \to X \in \mathscr{P}\left(\{i_1,\, i_3\}\right) \;:\; C_2^1 \text{ règles possibles,} \\ \quad \vdots \\ Y = \{i_{m-1},\, i_m\} \to X \in \mathscr{P}\left(\{i_{m-1},\, i_m\}\right) \;:\; C_2^1 \text{ règles possibles,} \end{cases}$$

au total $C_m^2\,(C_2^1)$ règles possibles.

Pour Y dans la partie de I à 3 éléments

Pour les C_m^3 Y possibles, on a :

$$\begin{cases} Y{=}\{i_1,\, i_2, i_3\} \to X \in \mathscr{P}\left(\{i_1,\, i_2, i_3\}\right) \;: & C_3^1 + C_3^2 \text{ règles possibles,} \\ Y{=}\{i_1,\, i_2, i_4\} \to X \in \mathscr{P}\left(\{i_1,\, i_2, i_4\}\right) \;: & C_3^1 + C_3^2 \text{ règles possibles,} \\ \quad \vdots & \\ Y{=}\{i_{m-2},\, i_{m-1},\, i_m\} \to X \in \mathscr{P}\left(\{i_{m-1},\, i_m\}\right) \;: & C_3^1 + C_3^2 \text{ règles possibles,} \end{cases}$$

au total $C_m^3\,(C_3^1 + C_3^2)$ règles possibles, et ainsi de suite.

Pour Y dans la partie de I à m éléments

Pour $Y = \{i_1,\, i_2, \ldots,\, i_m\}$, X prend sa valeur dans $\mathscr{P}\left(\{i_1,\, i_2, \ldots,\, i_m\}\right)$, on a donc :
$C_m^1 + C_2^2 + \ldots + C_m^{m-1}$ règles possibles.
En faisant la somme, on obtient :

$$\begin{aligned} S_m &= C_m^2\left(C_2^1\right) + C_m^3\left(C_3^1 + C_3^2\right) + \cdots + C_m^m\left(C_m^1 + C_m^2 + \cdots + C_m^{m-1}\right) \\ &= C_m^2\left(2^2 - 2\right) + C_m^3\left(2^3 - 2\right) + \cdots + C_m^m\left(2^m - 2\right) \\ &= C_m^2 2^2 + C_m^3 2^3 + \cdots + C_m^m 2^m - 2\left(C_m^2 + C_m^3 + \cdots + C_m^{m-1} + C_m^m\right). \end{aligned}$$

Remarquons que :

$$\begin{aligned} 3^m &= (2+1)^m \\ &= \sum_{k=0}^{m} C_m^k 2^k 1^{m-k} \\ &= \sum_{k=0}^{m} C_m^k 2^k 1 \\ &= C_m^0 2^0 + C_m^1 2^1 + \cdots + C_m^m 2^m. \end{aligned}$$

Donc, notre somme s'écrit :

$$\begin{aligned} S_m &= 3^m - 1 - 2m - 2\left(2^m - 1 - m\right) \\ &= 3^m - 1 - 2m - 2^{m+1} + 2 + 2m \\ &= 3^m - 2^{m+1} + 1. \end{aligned}$$

En conclusion, pour un ensemble d'attributs de taille m, le nombre des règles possibles est :

$$S_m = 3^m - 2^{m+1} + 1$$

Pour se donner une idée de l'ordre de grandeur, avec $m = 10$, on peut recenser jusqu'à 57002 règles possibles. supposons que seules 1% de ces règles soient valides, l'utilisateur va se retrouver alors devant 570 règles (dont la majorité sont redondantes) à interpréter. Or, dans la pratique, on a à faire à plusieurs dizaines, voir plusieurs centaines de variables. Face à ce nombre trop important des règles, l'utilisateur va se retrouver au point de départ, à savoir, la découverte des connaissances à partir des données volumineuses (cette fois, à partir des règles). C'est justement cette limite qui a motivé les recherches des ensembles minimaux des règles plus informatives et non redondantes que l'on appelle communément « Bases des règles ». Plus précisément, au lieu d'énumérer l'ensemble de toutes les règles valides, on se contente d'extraire un ensemble minimal composé par les règles les plus informatives et à partir desquelles on peut dériver (dans le sens de retrouver), si l'on souhaite, l'ensemble des règles valides via ce qu'on appelle *les axiomes d'inférence*. Nous tenons à souligner que la minimalité de cet ensemble des règles est relative aux axiomes d'inférence considérés. Pour s'assurer qu'aucune information utile ne soit perdue par cette réduction, elle doit se faire en respectant les conditions de *dérivabilité et d'informativité*. On peut retrouver ces deux concepts, entre autres, dans les travaux de (Gasmi *et al.*, 2006 ; Pasquier, 2000a ; Hamrouni *et al.*, 2011). Le terme **dérivabilité** est utilisé pour spécifier que les axiomes d'inférence sont valides ou corrects (selon les auteurs), c'est-à-dire qu'ils ne permettent de dériver que des règles valides et qu'ils sont complets (ils permettent de retrouver l'ensemble de toutes les règles valides). De son coté, **informativité** signifie qu'à partir d'une règle et en utilisant les axiomes d'inférence, on peut toujours retrouver les valeurs des mesures (en occurrence le Support et la Confiance) des règles dérivées.

3.3.1 Bases des règles Support-Confiance valides

D'abord, rappelons qu'une règle valide selon les mesures Support et Confiance est une règle $X \rightarrow Y$ telle que $\text{Supp}(X \rightarrow Y) \geqslant minSupp$ et $\text{Conf}(X \rightarrow Y) \geqslant minConf$, où $minSupp$ et $minConf$ désignent respectivement les valeurs minimales de Support et celle de confiance qui sont fixées par l'utilisateur. Le Support quantifie la proportion d'objets ou de transactions contenant simultanément les deux motifs X et Y (prémisse et conséquent) par rapport au nombre total d'objets du contexte. La confiance désigne en fait, la probabilité conditionnelle d'apparition de motif Y sachant que X est dans la transaction. Dans les définitions des bases des règles d'association, on distingue souvent les bases des règles exactes et celles des règles approximatives. Nous avons donné la signification d'une règle exacte et approximative dans l'approche de R. Gras (chapitre 2, § 2.4). Nous avons vu qu'une règle $X \rightarrow Y$ est exacte si elle ne souffre d'aucun contre-exemple ($n_{X\overline{Y}} = 0$). Cette même description se traduit par $\text{Conf}(X \rightarrow Y) = 1$ dans le cadre d'utilisation de la mesure Confiance. En effet, pour tout X, Y de $\mathscr{P}(\mathcal{I})$ d'un contexte binaire $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$, $n_{XY} + n_{X\overline{Y}}$ est toujours égal à n_X, donc si $n_{X\overline{Y}} = 0$, alors, $n_{XY} = n_X$ et $\text{Conf}(X \rightarrow Y) = \frac{n_{XY}}{n_X} = \frac{n_x}{n_X} = 1$. Rappelons que pour tout motif X de $\mathscr{P}(\mathcal{I}), n_X$ désigne le nombre d'objets ou de transactions contenant X ($n_X = \text{Card}(\phi(X))$ où ϕ désigne l'opérateur d'extension défini sur $\mathscr{P}(\mathcal{I})$). Dans le cas contraire, i. e. $\text{Conf}(X \rightarrow Y)$ est strictement plus petite que 1, la règle $X \rightarrow Y$ est appelée

règle approximative. Nous allons maintenant présenter quelques bases des règles Support-Confiance valides.

Base de Duquenne-Guigues

La base de Duquenne-Guigues définie par Guigues et Duquenne en 1986 et redéfinie dans Pasquier (2000a) pour être compatible aux concepts d'extraction des bases des règles d'association est une base des règles exactes définie à partir des motifs pseudo-fermés (définition 3.5). La dérivation des autres règles exactes est assurée par les axiomes d'inférence d'Armstrong donnés ci-dessous :

Si $X \supseteq Y$, alors $X \rightarrow Y$;
Si $X \rightarrow Y$ et $Y \rightarrow Z$, alors $X \rightarrow Z$ (transitivité) ;
Si $X \rightarrow Y$ et $Z \rightarrow T$, alors $X \cup Z \rightarrow Y \cup T$.

Définition 3.6. *Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte binaire d'extraction. Désignons par PF l'ensemble des motifs pseudo-fermés de ce contexte. La base de Duquenne-Guigues est l'ensemble des règles de type $X \rightarrow (\gamma(X) \setminus X)$ avec X un pseudo-fermé.*

$$BDG = \{r : X \rightarrow (\gamma(X) \setminus X) \text{ avec } X \in PF, X \neq \emptyset\} \tag{3.6}$$

Soulignons qu'il est possible d'étendre ces axiomes d'inférence. On peut citer par exemple :

Augmentation à gauche : Si $X \rightarrow Y$ est exacte, alors pour tout motif Z, $X \cup Z \rightarrow Y$ l'est aussi ;
Addition à gauche : Si $X \rightarrow Y$ et $Z \rightarrow Y$, sont exactes, alors $X \cup Z \rightarrow Y$ l'est aussi ;
Addition à droite : Si $X \rightarrow Y$ et $X \rightarrow Z$, sont exactes, alors $X \rightarrow Y \cup Z$ l'est aussi.

Toutes les règles exactes peuvent être déduites de cette base, mais les règles dérivées ne sont pas toujours valides et les axiomes d'inférence d'Armstrong ne permettent pas de retrouver la précision (les mesures) des règles dérivées. En effet, supposons que $X \rightarrow Y$ soit un élément de BDG et utilisons « l'augmentation à gauche » des axiomes d'Armstrong pour dériver une règle, situation illustrée par la figure 3.1.
Si $\text{Conf}(X \rightarrow Y) = 1$, alors $\forall Z \in \mathscr{P}(\mathcal{I})$, $\text{Conf}(X \cup Z \rightarrow Y) = \frac{n_{XZY}}{n_{XZ}} = \frac{n_{XZ}}{n_{XZ}} = 1$.

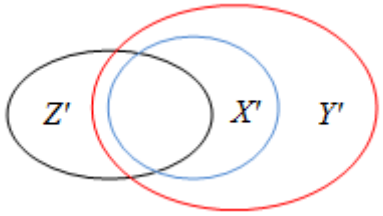


FIGURE 3.1 – Augmentation à gauche : Axiome d'Armstrong

Par contre, pour le cas de Support, on a :

$$\begin{aligned} \text{Supp}(X \cup Z \rightarrow Y) &= \frac{n_{XZY}}{n} = \frac{n_{XZ}}{n}, \\ \text{Supp}(X \rightarrow Y) &= \frac{n_{XY}}{n} = \frac{n_X}{n}, \\ \text{donc, Supp}(X \cup Z \rightarrow Y) &\leqslant \text{Supp}(X \rightarrow Y). \end{aligned}$$

Ainsi, même si le support de la règle $X \rightarrow Y$ soit plus grand que le seuil *minSupp*, rien ne garantit que $\text{Supp}(X \cup Z \rightarrow Y)$ va rester plus grand que *minSupp* (étant donné qu'il est plus petit que $\text{Supp}(X \rightarrow Y)$). Autrement dit, non seulement, on ne peut pas prévoir le support des prémisses des règles dérivées, mais on peut aussi avoir une règle non valide puisque le support $\text{Supp}(X \cup Z \rightarrow Y)$ peut descendre en dessous de *minSupp*.

Exemple 4 (Exemple de base de Duquenne-Guigues)**.**

Considérons le contexte binaire décrit dans le tableau 2.1, cherchons l'ensemble des pseudo-fermés de ce contexte d'extraction. Soulignons d'abord que si $\emptyset$ est un pseudo-fermé (donc il n'est pas fermé), alors il existe un motif l tel que $\gamma(\emptyset) = l$. Comme l'ensemble vide est un sous-ensemble de tous les ensembles, un pseudo-fermé doit nécessairement contenir le motif l parce qu'un pseudo-fermé doit contenir les fermetures de tous ses sous-ensembles qui sont des pseudo-fermés. Dans notre cas (contexte $\mathcal{K}$ du tableau 2.1), $\emptyset$ est fermé ($\gamma(\emptyset) = \emptyset$), par conséquent, tous les motifs de taille 1, c'est-à-dire les 1-itemset (un k-itemset est un motif formé par k items) qui ne sont pas des fermés sont des pseudo-fermés. Donc, C, E et F sont des pseudo-fermés. À partir de ces pseudo-fermés, on peut tester progressivement les autres motifs. En prenant un $minSupp = 1/3$, la base de Duquenne-Guigues est donnée dans le tableau 3.2.

Pseudo-fermés	Fermetures	Règles exactes	Supports
E	BE	$E \rightarrow B$	2/3
C	ABC	$C \rightarrow AB$	1/2
F	$ABEF$	$F \rightarrow ABE$	1/3

Tableau 3.2 – Base de Duquenne-Guigues

Bases de Luxenburger

Toujours dans le cadre de réduction du nombre de règles extraites en même temps pour s'assurer qu'aucune information pertinente ne soit perdue, Luxenburger a défini une famille des bases des règles approximatives (règles de confiance strictement plus petite que 1). En effet, comme nous l'avons vu lors de l'étude portant sur l'intensité d'implication, ce n'est pas parce qu'on a quelques contre-exemples qu'il faut rejeter une règle, il est très fréquent d'avoir des règles valides ayant des contre-exemples. La question est juste de savoir décider jusqu'où les contres exemples sont acceptables ; d'où l'importance de la génération des règles approximatives. Dans ce paragraphe, nous allons voir deux bases des règles approximatives : base propre et base de couverture. Nous allons d'abord voir quelques propriétés sur lesquelles sont fondées ces bases des règles.

Propriété 3.7. *Soient X et Y deux motifs fermés d'un contexte binaire $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$.*

1. *Si $X \subset Y$, alors $Conf(X \rightarrow Y \setminus X) < 1$.*
2. *$\forall X, Y, Z \in \mathscr{P}(\mathcal{I})$, si $X \subset Y \subset Z$ alors :*
   $Conf(X \rightarrow Y \backslash X) \times Conf(Y \rightarrow Z \backslash Y) = Conf(X \rightarrow Z \backslash X)$.

Preuve.
Désignons par (ψ, ϕ) une correspondance de Galois associée au contexte binaire $\mathcal{K}$. En utilisant l'antitonie de ϕ, à partie de $X \subset Y$, on obtient $\phi(X) \supseteq \phi(Y)$; ce qui signifie que, soit $\phi(X) \supset \phi(Y)$, soit $\phi(X) = \phi(Y)$. Si $\phi(X) = \phi(Y)$, alors on a nécessairement $\psi(\phi(X)) = \gamma(X) = \psi(\phi(Y)) = \gamma(Y)$, une situation impossible dans notre cas puisque X et Y sont fermés (un fermé X est unique dans sa classe[3] que l'on note par $[X]$). Donc, on est forcément dans le deuxième cas, c'est-à-dire : $\phi(X) \supset \phi(Y)$; ce qui nous permet de déduire que $|\phi(X)| > |\phi(Y)|$, donc $\text{Supp}(X) > \text{Supp}(Y)$. Calculons maintenant la valeur de la mesure confiance de la règle $(X \to Y \setminus X)$.

$$\begin{aligned} \text{Conf}(X \to Y \setminus X) &= \frac{\text{Supp}(X \cup Y \setminus X)}{\text{Supp}(X)} \\ &= \frac{\text{Supp}(X \cup Y)}{\text{Supp}(X)} \\ &= \frac{\text{Supp}(Y)}{\text{Supp}(X)}, \\ \text{donc } \text{Conf}(X \to Y \setminus X) &< 1. \end{aligned}$$

La base des règles approximatives va faire intervenir des motifs (prémisses et conséquents) dans des classes différentes. Pour le second point, on a successivement :

$$\begin{aligned} \text{Conf}(X \to Y \setminus X) &= \frac{\text{Supp}(X \cup Y)}{\text{Supp}(X)} = \frac{\text{Supp}(Y)}{\text{Supp}(X)}; \\ \text{Conf}(Y \to Z \setminus Y) &= \frac{\text{Supp}(Z \cup Y)}{\text{Supp}(Y)} = \frac{\text{Supp}(Z)}{\text{Supp}(Y)}; \\ \text{Conf}(X \to Y \setminus X) \times \text{Conf}(Y \to Z \setminus Y) &= \frac{\text{Supp}(Z)}{\text{Supp}(X)}; \\ \text{Conf}(X \to Z \setminus X) &= \frac{\text{Supp}(Z)}{\text{Supp}(X)}. \end{aligned}$$

□

Propriété 3.8 (Inférence de règles basée sur la fermeture : $\mathcal{IRF}$)**.**
Dans un contexte binaire binaire $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$, désignons par γ l'opérateur de fermeture de connexion de Galois. Pour tous motifs $X, Y \in \mathscr{P}(\mathcal{I})$ tels que $X \subset Y$, on a :

$$\begin{aligned} \mathit{Supp}(X \to Y \setminus X) &= \mathit{Supp}(\gamma(Y)), \\ \mathit{Conf}(X \to Y \setminus X) &= \frac{\mathit{Supp}(\gamma(Y))}{\mathit{Supp}(\gamma(X))}, \\ \text{et } X \to Y \setminus X \text{ est valide} &\Leftrightarrow \gamma(X) \to \gamma(Y) \setminus \gamma(X) \text{ est valide.} \end{aligned}$$

3. La classe de X, notée par $[X]$ est définie par $[X] = \{Y \in \mathscr{P}(\mathcal{I}) / \gamma(X) = \gamma(Y)\}$.

Preuve.

$$\begin{aligned}
\text{Supp}(X \rightarrow Y \setminus X) &= \text{Supp}(X \cup Y) \\
&= \text{Supp}(Y) \\
&= \text{Supp}(\gamma(Y)) \\
\text{Conf}(X \rightarrow Y \setminus X) &= \frac{\text{Supp}(X \cup Y)}{\text{Supp}(X)} \\
&= \frac{\text{Supp}(Y)}{\text{Supp}(X)} \\
&= \frac{\text{Supp}(\gamma(Y))}{\text{Supp}(\gamma(X))}
\end{aligned}$$

On peut se servir de ces deux premiers résultats pour montrer que :

$$\begin{cases} \text{Supp}(\gamma(X) \rightarrow \gamma(Y) \setminus \gamma(X)) = \text{Supp}(X \rightarrow Y \setminus X), \\ \text{Conf}(\gamma(X) \rightarrow \gamma(Y) \setminus \gamma(X)) = \text{Conf}(X \rightarrow Y \setminus X), \end{cases}$$

d'où l'équivalence entre la validité des règles $(X \rightarrow Y \setminus X)$ et $(\gamma(X) \rightarrow \gamma(Y) \setminus \gamma(X))$. □

Base propre

Considérons un contexte binaire $\mathcal{K}$ et désignons par $\mathcal{F}$ l'ensemble des motifs fermés fréquents. La base propre pour les règles d'association approximatives est définie par :

$$BP = \{r : F_1 \rightarrow F_2 \setminus F_1 \ \textit{tels que}\ F_1, F_2 \in \mathcal{F}, F_1 \subset F_2 \text{ et } \text{Conf}(r) \geqslant \textit{minConf}\} \qquad (3.7)$$

La propriété 3.7 nous assure que pour une règle r de la base propre, $\text{Conf}(r) < 1$. La dérivation des autres règles approximatives s'effectuera grâce aux axiomes d'inférence basée sur la fermeture.

Exemple 5. *Pour trouver la base propre d'un contexte quelconque, il faut d'abord extraire les motifs fermés. Plusieurs algorithmes ont été proposé pour extraire les motifs fermés, en l'occurrence les algorithmes Close, A-Close, Close*$^+$ *(Pasquier et al., 1999a,b,c). Pour le contexte binaire $\mathcal{K}$ du tableau 2.1, la figure 3.2 présente le treillis des motifs fermés avec leurs supports absolus.*

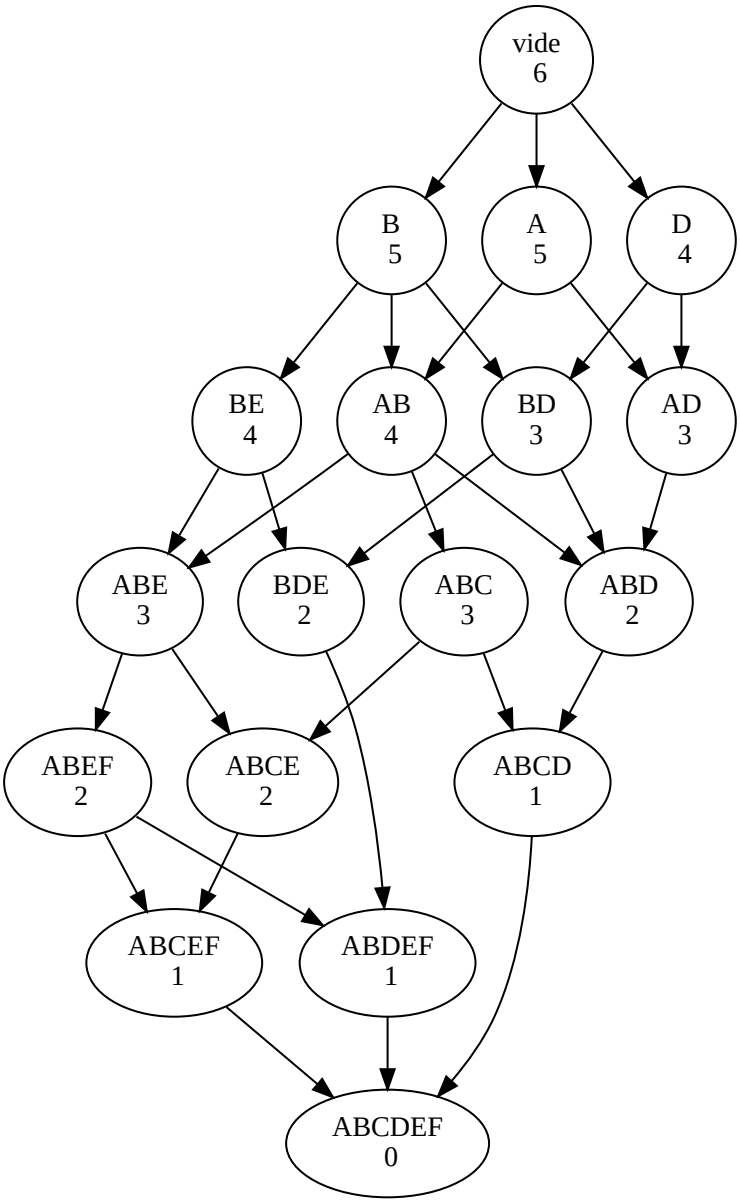


FIGURE 3.2 – Treillis des motifs fermés

À partir de ce treillis, on peut déduire les motifs fermés fréquents et par la suite, la base propre. La base propre du contexte décrit dans le tableau 2.1, pour un $minSupp = 1/2$ *et* $minConf = 0,6$ *est donnée dans le tableau 3.3.*

Fermés	Sur-ensembles fermés fréquents	Base propre	Supports	Confiances
B	BE	$B \rightarrow E$	2/3	4/5
	AB	$B \rightarrow A$	2/3	4/5
	BD	$B \rightarrow D$	1/2	3/5
	ABE	$B \rightarrow AE$	1/2	3/5
	ABC	$B \rightarrow AC$	1/2	3/5
A	AB	$A \rightarrow B$	2/3	4/5
	AD	$A \rightarrow D$	1/2	3/5
	ABE	$A \rightarrow BE$	1/2	3/5
	ABC	$A \rightarrow BC$	1/2	3/5
D	BD	$D \rightarrow B$	1/2	3/4
	AD	$D \rightarrow A$	1/2	3/4
BE	ABE	$BE \rightarrow A$	1/2	3/4
AB	ABE	$AB \rightarrow E$	1/2	3/4
	ABC	$AB \rightarrow C$	1/2	3/4
BD				
AD				
ABE				
ABC				

Tableau 3.3 – Exemple de base propre

On peut extraire les autres règles approximatives avec leurs mesures (support et Confiance en utilisant la base propre et les axiomes $\mathcal{IRF}$.

Base de couverture

Avant de définir la base de couverture, nous allons d'abord voir la notion de couverture d'un élément dans un ensemble ordonné.

Définition 3.7. *Soit $(E, <)$ un ensemble ordonné. La relation de couverture sur E, notée par $\prec$ est définie par : $\forall X, Y \in E, X \prec Y$ si $X < Y$ et $\nexists Z \in E, X < Z < Y$.*
Si $X \prec Y$, alors on dit que X est couvert par Y ou encore que Y couvre X.

Considérons maintenant trois motifs fermés F, G et H tels que $F \subset G \subset H$. Selon la propriété de transitivité de Confiance : Conf($F \rightarrow G \setminus F$)×Conf($G \rightarrow H \setminus G$)=Conf($F \rightarrow H \setminus F$). Donc connaissant Conf($F \rightarrow G \setminus F$) et Conf($G \rightarrow H \setminus G$), on peut déduire Conf($F \rightarrow H \setminus F$). De plus, Conf($F \rightarrow H \setminus F$) est toujours plus petite que le maximum entre Conf($F \rightarrow G \setminus F$) et Conf($G \rightarrow H \setminus G$). Or, selon l'approche d'extraction utilisant la mesure Confiance, l'importance d'une règle croît avec la valeur de sa Confiance. Selon cette logique, en 1991, Luxenburger a défini la base de couverture, une base formée par des règles de type $X \rightarrow Y \setminus X$, avec $X \prec Y$. La base de couverture est définie par :

$$BC = \{r : F_1 \rightarrow F_2 \setminus F_1 \text{ tels que } F_1, F_2 \in \mathcal{F}, F_1 \prec F_2 \text{ et } \text{Conf}(r) \geqslant minConf\}. \quad (3.8)$$

On peut remarquer que la base de couverture est un sous-ensemble de la base propre. En fait, elle correspond à la réduction transitive de la base propre. C'est à dire qu'on peut l'avoir en

supprimant les arcs transitifs du graphe représentant la base propre. Pour le contexte binaire $\mathcal{K}$ du tableau 2.1, voici la base de couverture :

$$\begin{aligned} BC \quad = \quad & \{B \rightarrow E, B \rightarrow A, B \rightarrow D, A \rightarrow B, A \rightarrow D, \\ & D \rightarrow B, D \rightarrow A, BE \rightarrow A, AB \rightarrow E, AB \rightarrow C\}. \end{aligned}$$

Selon (Pasquier, 2000a), la base de Duquenne-Guigues et les bases de Luxenburger (base propre et base de couverture) ne sont pas des bases les plus informatives pour l'utilisateur. En effet, entre deux règles r et r' de même Support et même Confiance, la plus informative est celle qui a un conséquent maximal et une prémisse minimale. La base de Duquenne-Guigues et les bases de Luxenburger sont formées par des prémisses pseudo-fermées ou fermées et il est clair que ces motifs ne sont pas minimaux (au sens de l'inclusion) dans leur classe (Classe de X, notée $[X]$ désigne l'ensemble des motifs ayant la même fermeture que X). De plus, comme nous l'avons fait remarquer plus haut, l'application des axiomes d'Armstrong ne permet pas de retrouver le support des motifs des règles dérivées. Ainsi, pour améliorer la base de Duquenne-Guigues pour les règles exactes et les bases de Luxenberger (base propre et base de couverture) pour les règles approximatives, trois nouvelles bases sont proposées dans (Pasquier, 2000a). Avoir des règles non redondantes selon la définition 3.8 est l'une des idées maîtresses utilisées pour la construction de ces nouvelles bases des règles.

Définition 3.8 (Règle non redondante minimale)**.** *Soit $r_1 : X_1 \rightarrow Y_1$ une règle valide. La règle r_1 est non redondante minimale s'il n'existe pas une règle valide $r_2 : X_2 \rightarrow Y_2$ telle que Supp(r_1)=Supp(r_2), Conf(r_1)=Conf(r_2), $X_2 \subset X_1$ et $Y_2 \supset Y_1$.*

Base générique des règles exactes

Par définition, les générateurs d'un motif fermé X sont les éléments incomparables les plus petits (au sens de l'inclusion) constituants la classe de X, ainsi, au lieu de prendre un pseudo-fermé, pour avoir une prémisse minimale, on peut plutôt prendre un générateur. Si on désigne par $\mathcal{F}$ l'ensemble des fermés fréquents pour un seuil *minSupp* fixé et par $\mathcal{G}_Y$ l'ensemble des générateurs d'un motif fermé Y, alors la base générique BG est définie par :

$$BG = \{r : X \rightarrow Y \setminus X \text{ telle que } X \in \mathcal{G}_Y, Y \in \mathcal{F} \text{ et } X \neq Y\}. \tag{3.9}$$

Exemple 6. *Considérons le contexte binaire $\mathcal{K}$ du tableau 2.1, en fixant un minSupp = 1/3, la base générique qu'on peut extraire de ce contexte est donnée dans le tableau 3.4. On ne peut tirer aucune règle avec les autres générateurs (A, B, D, AB, BD, AD) même s'ils sont fréquents.*

Les règles de BG sont des règles intra-classes, c'est-à-dire, des règles entre prémisse et conséquent d'une même classe. On a montré dans (Pasquier, 2000b) que cette nouvelle base des règles exactes contient beaucoup plus de règles que la base de Duquenne-Guiges (DG). Sauf que les règles dans BG sont beaucoup plus informatives que les règles dans DG parce dans BG, les règles sont de prémisse minimale et de conséquent maximal. De plus, BG est une base des règles exactes valides, c'est-à-dire, toutes les règles valides ainsi que leurs précisions (Support et Confiance) peuvent être déduites de BG.

Générateurs	Fermetures	Base Générique			Support
E	BE	E	$\rightarrow$	B	2/3
AE	ABE	AE	$\rightarrow$	B	1/2
DE	BDE	DE	$\rightarrow$	B	1/3
C	ABC	C	$\rightarrow$	B	1/2
F	$ABEF$	F	$\rightarrow$	ABE	1/3
CE	$ABCE$	CE	$\rightarrow$	AB	1/3

Tableau 3.4 – Base générique des règles exactes

Remarque Examinons attentivement les deux règles $r : E \rightarrow B$ et $r' : AE \rightarrow B$. D'abord ces deux règles sont des règles exactes (Conf($E \rightarrow B$) = Conf($AE \rightarrow B$)=1). Ensuite, elles ont les mêmes conséquents et des prémisses comparables. Logiquement, r est beaucoup plus informative que r'. En effet, selon r, si E est présent dans une transaction (un objet), alors on y trouve toujours l'item B. Or, selon r', la présence de B dans une transaction est conditionnée par la présence simultanée des items A et E. Comme les deux règles sont exactes, on peut affirmer que le minimum des conditions suffisantes à la présence de B est tout simplement la présence de E. Sauf que, les deux règles ont des Supports différents, concrètement cela signifie que les proportions des transactions contenants les items composants la prémisse et le conséquent de ces deux règles sont différents. Mais, si l'utilisateur se fixe un *minSupp* (1/3 par exemple), c'est qu'il est prêt à prendre en considération toutes les règles confiance-valides de support plus grand que *minSupp*. Sachant que deux règles sont exactes, si on considère le *minSupp* comme l'unique balise de validation, c'est-à-dire qu'on ne tient pas en compte des valeurs des supports lorsqu'on sait qu'elles dépassent le *minSupp*, alors la règle r' n'apporterait pas d'information supplémentaire par rapport à r. Donc, selon cette logique, r' est une règle redondante à l'égard de la règle r. Ce raisonnement est utilisé dans (Gasmi *et al.*, 2006) pour définir une nouvelle base générique que nous allons voir ultérieurement. Le même raisonnement nous servira pour définir une semi-base des règles M_{GK}-valides.

Base informative pour les règles approximatives

Selon la propriété 3.7, pour deux motifs fermés F_1 et F_2 tels que $F_1 \subset F_2$, Conf($F_1 \rightarrow F_2 \setminus F_1$) est toujours plus petit que 1. Qu'en est il de la Conf($X \rightarrow Y \setminus X$) lorsque $X \in [F_1]$ et $Y \in [F_2]$? Nous avons la propriété 3.9 qui donne un élément de réponse à cette question.

Propriété 3.9. *Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte binaire d'extraction, F_1 et F_2 deux motifs fermés de $\mathcal{K}$ tels que $F_1 \subset F_2$. Pour tous motifs X, Y de $\mathscr{P}(\mathcal{I})$ tels que $X \in [F_1]$ et $Y \in [F_2]$, nous avons les relations suivantes :*

1. *Supp($X \rightarrow Y \setminus X$) = Supp($F_1 \cup F_2$) = Supp($F_2$),*
2. *Conf($X \rightarrow Y \setminus X$) = Conf($F_1 \rightarrow F_2 \setminus F_1$),*
3. *Conf($X \rightarrow Y \setminus X$) < 1.*

Preuve.

$$
\begin{aligned}
\mathrm{Supp}(X \to Y \setminus X) &= \mathrm{Supp}(X \cup Y \setminus X) &&= \mathrm{Supp}(X \cup Y) \\
&= \mathrm{Supp}(\gamma(X \cup Y)) &&= \mathrm{Supp}(\gamma(\gamma(X) \cup \gamma(Y))) \\
&= \mathrm{Supp}(\gamma(F_1 \cup F_2)) &&= \mathrm{Supp}(F_1 \cup F_2) \\
&= \mathrm{Supp}(F_2). &&
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Conf}(X \to Y \setminus X) &= \frac{\mathrm{Supp}(X \cup Y \setminus X)}{\mathrm{Supp}(X)} &&= \frac{\mathrm{Supp}(X \cup Y)}{\mathrm{Supp}(X)} \\
&= \frac{\mathrm{Supp}(\gamma(X \cup Y))}{\mathrm{Supp}(X)} &&= \frac{\mathrm{Supp}(\gamma(\gamma(X) \cup \gamma(Y)))}{\mathrm{Supp}(\gamma(X))} \\
&= \frac{\mathrm{Supp}(\gamma(F_1 \cup F_2))}{\mathrm{Supp}(F_1)} &&= \frac{\mathrm{Supp}(F_1 \cup F_2)}{\mathrm{Supp}(F_1)} \\
&= \mathrm{Conf}(F_1 \to F_2 \setminus F_1) &&< 1 \text{ (selon la propriété 3.7).}
\end{aligned}
$$

□

La propriété 3.9 nous permet d'affirmer que d'une part, une règle inter-classe (entre deux classes différentes de fermetures comparables) est toujours une règle approximative et, d'autre part, toutes les règles issues de deux classes différentes ont les mêmes Supports et mêmes Confiances. Donc, on peut déduire toutes les règles entre deux classes de fermetures comparables à partir de n'importe quelle règle entre ces mêmes classes. On peut donc choisir n'importe quelle règle $r : X \to Y \setminus Y$ pour représenter l'ensemble des règles liant [X] et [Y]. Nous avons vu précédemment qu'une règle est plus informative si elle a une prémisse minimale et un conséquent maximal. Comme on peut choisir n'importe quelle règle, pourquoi ne pas choisir la plus informative. C'est ainsi que (Pasquier, 2000a) a défini une nouvelle base des règles approximatives qui fait intervenir les motifs fermés et les générateurs. En désignant par $\mathcal{F}$ l'ensemble des fermés fréquents du contexte étudié et par $\mathcal{G}$ celui des générateurs, cette nouvelle base des règles approximatives est définie par :

$$BI = \{r : X \to Y \setminus X \text{ telle que } Y \in \mathcal{F}, X \in \mathcal{G}, \gamma(X) \subset Y \text{ et } \mathrm{Conf}(r) \geqslant minConf\}. \quad (3.10)$$

Cette nouvelle base des règles approximative est valide (Pasquier, 2000a), c'est-à-dire, toutes les règles valides approximatives peuvent être déduites avec leur Support et Confiance.

Exemple 7. *Pour le contexte binaire $\mathcal{K}$ du tableau 2.1 et en prenant un $minSupp = 0,5$ et $0,6$ comme $minConf$, la base informative est donnée dans le tableau 3.5.*

Générateurs	Fermetures	Sur-ensembles fermés fréquents	Base Informative	Supports	Confiance
A	A	AB	$A \rightarrow B$	2/3	4/5
		ABC	$A \rightarrow BC$	1/2	3/5
		AD	$A \rightarrow D$	1/2	3/5
		ABE	$A \rightarrow BE$	1/2	3/5
B	B	BE	$B \rightarrow E$	2/3	4/5
		ABE	$B \rightarrow AE$	1/2	3/5
		AB	$B \rightarrow A$	2/3	4/5
		ABC	$B \rightarrow AC$	1/2	3/5
		BD	$B \rightarrow D$	1/2	3/5
D	D	BD	$D \rightarrow B$	1/2	3/4
		AD	$D \rightarrow A$	1/2	3/4
E	BE	ABE	$E \rightarrow AB$	1/2	3/4
AB	AB	ABE	$AB \rightarrow E$	1/2	3/4
		ABC	$AB \rightarrow C$	1/2	3/4

Tableau 3.5 – Base informative des règles approximatives

Remarque 1. *Étant donné un motif fermé X, la classe $[X]$ peut contenir plus d'un générateur ; par contre le fermé est unique. Par conséquent, la taille de la base propre est plus petite que celle de la base informative. D'un côté, l'utilisation de la base informative a augmenté le nombre des règles extraites, mais d'un autre côté, la base BI est constituée des règles beaucoup plus informatives par rapport à la base propre. Autrement dit, on a perdu en quantité mais on a gagné en qualité. Pour le cas particulier du contexte 2.1, la base propre du tableau 3.3 et la base informative du tableau 3.5 sont très similaires, tout simplement parce que dans ce contexte, soit les générateurs sont fermés, soit ils ont des fermetures non fréquentes.*

Réduction transitive de la base informative

Comme dans la réduction transitive de la base propre qui nous a donné la base de couverture, on peut utiliser le même raisonnement avec la base informative des règles approximatives. Au lieu d'évaluer une règle entre un générateur G et un fermé F tels que $\gamma(G) \subset F$, on peut seulement se contenter de la règle $G \rightarrow F \setminus G$ avec $\gamma(G) \prec F$ comme dans la figure 3.3 (où $\prec$ désigne l'opérateur de couverture décrit dans la définition 3.7). En effet, considérons trois fermés F_1, F_2, F_3 tels que $F_1 \subset F_2 \subset F_3$ et leurs générateurs respectifs G_1, G_2 et G_3. Montrons qu'on peut exprimer les mesures (Support et Confiance) de la règle $r_3 : G_1 \rightarrow F_3 \setminus G_1$ en fonction de celles de $r_1 : G_1 \rightarrow F_2 \setminus G_1$ et $r_2 : G_2 \rightarrow F_3 \setminus G_2$.

$$\begin{array}{ccccc} \text{Supp}(G_1 \rightarrow F_3 \setminus G_1) & = & \text{Supp}(G_1 \cup F_3) & = & \text{Supp}(F_3) \\ \text{Supp}(G_2 \rightarrow F_3 \setminus G_2) & = & \text{Supp}(G_2 \cup F_3) & = & \text{Supp}(F_3) \end{array}$$

Donc $\text{Supp}(G_1 \rightarrow F_3 \setminus G_1) = \text{Supp}(G_2 \rightarrow F_3 \setminus G_2)$. Selon cette égalité, connaissant le support de $G_2 \rightarrow F_3 \setminus G_2$, avec $\gamma(G_2) \prec F_3$, on peut connaitre le support de toutes les règles $r : G \rightarrow F_3 \setminus G$ telles que $\gamma(G) \subset F_3$. Examinons maintenant la Confiance de r_3, r_2 et r_1.

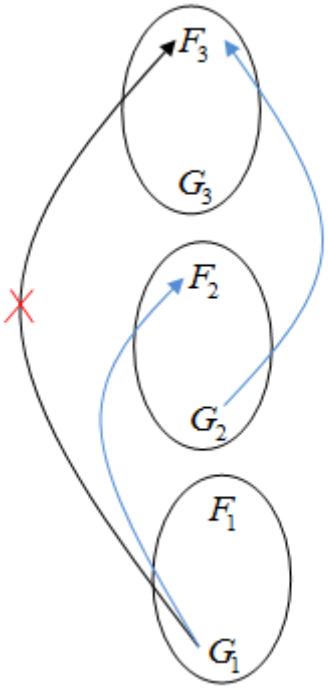


FIGURE 3.3 – Réduction transitive

$$\begin{aligned} \text{Conf}(G_1 \rightarrow F_2 \setminus G_1) &= \frac{\text{Supp}(G_1 \cup F_2)}{\text{Supp}(G_1)} = \frac{\text{Supp}(F_2)}{\text{Supp}(G_1)} \\ \text{Conf}(G_2 \rightarrow F_3 \setminus G_2) &= \frac{\text{Supp}(G_2 \cup F_3)}{\text{Supp}(G_2)} = \frac{\text{Supp}(F_3)}{\text{Supp}(G_2)} \end{aligned}$$

En faisant le produit, on a (transitivité de Confiance) :

$$\text{Conf}(G_1 \rightarrow F_2 \setminus G_1) \times \text{Conf}(G_2 \rightarrow F_3 \setminus G_2) = \frac{\text{Supp}(F_3)}{\text{Supp}(G_1)}.$$

Connaissant le Support et la Confiance de $r_1 : G_1 \rightarrow F_2 \setminus G_1$ et $r_2 : G_2 \rightarrow F_3 \setminus G_2$, on peut déduire ceux de la règle $r_3 : G_1 \rightarrow F_3 \setminus G_1$. De plus, Conf($r_3$) est toujours plus petit que le maximum entre Conf(r_1) et Conf(r_2). En désignant par $\mathcal{F}$ l'ensemble des fermés et par $\mathcal{G}$ l'ensemble des générateurs, les propriétés sur la transitivité de la mesure Confiance ont permis de définir la réduction transitive de la base informative pour les règles d'association approximatives donnée ci-après :

$$RTI = \{r : G \rightarrow F \setminus G \text{ telle que } F \in \mathcal{F}, G \in \mathcal{G}, \gamma(G) \prec F \text{ et } \text{Conf}(r) \geqslant minConf\}. \quad (3.11)$$

Exemple 8. *En reprenant le contexte binaire $\mathcal{K}$ du tableau 2.1, nous avons dans le tableau 3.6, avec $minSupp = 1/2$ et $minConf = 0,6$. Soulignons que comme son nom l'indique, la réduction transitive de la base informative n'est autre qu'un sous-ensemble obtenu en supprimant les règles entre prémisse et conséquent qui ne sont pas liés par l'opérateur de couverture.*

Générateurs	Fermetures	Sur-ensembles fermés fréquents	RTI	Supports	Confiance
A	A	AB	$A \rightarrow B$	2/3	4/5
		AD	$A \rightarrow D$	1/2	3/5
B	B	BE	$B \rightarrow E$	2/3	4/5
		AB	$B \rightarrow A$	2/3	4/5
		BD	$B \rightarrow D$	1/2	3/5
D	D	BD	$D \rightarrow B$	1/2	3/4
		AD	$D \rightarrow A$	1/2	3/4
E	BE	ABE	$E \rightarrow AB$	1/2	3/4
AB	AB	ABE	$AB \rightarrow E$	1/2	3/4
		ABC	$AB \rightarrow C$	1/2	3/4

Tableau 3.6 – Réduction transitive de la base informative

Nouvelle base générique (Gasmi *et al.*, 2006)

La base générique des règles exactes (BG) et la base informative des règles approximatives (BI) sont des bases des règles extraites sans perte d'information (on peut retrouver l'ensemble des règles valides ainsi que leurs Supports et Confiances). Toutefois, selon (Gasmi *et al.*, 2006), ces deux bases génèrent encore beaucoup trop de règles. Pasquier a proposé la réduction transitive pour justement pallier le problème de sur-abondance des règles dans BI. Néanmoins, selon toujours Gasmi *et al.*, la taille de cette base reste encore trop importante. De plus, la construction de la base générique est faite de manière syntaxique, sans accorder une importance à la dimension sémantique des règles. À titre d'exemple, prenons un cas dans la base générique des règles exactes (BG) et un autre cas dans la réduction transitive de la base des règles Informatives (RTI). Dans le cas des règles exactes, intéressons-nous aux règles $r_1 : E \rightarrow B$ et $r_2 : AE \rightarrow B$. Toutes les deux ont le même Support et même Confiance. Selon r_2, à chaque fois que A et E se trouvent dans une transaction, on observe aussi l'item B dans cette même transaction. Autrement dit, la présence simultanée des items A et E constitue une condition nécessaire à la présence de l'item B. Une conclusion que l'on peut tirer de cette règle, dans le cas d'une transaction de vente par exemple, est que, pour stimuler l'achat de l'article B, il faut d'abord favoriser l'achat de A et E. En conséquence, le gérant peut prendre la décision de faire une réduction sur les deux articles A et E. Pourtant selon r_1, il a juste fallu faire une réduction sur E pour activer l'achat de B. Pour ce gérant, la règle r_2 n'apporte aucune information utile du moment qu'il a la règle r_1. Dans un cas où A, B et E désignent des erreurs d'un élève, la règle r_2 veut faire croire que les deux erreurs A et E sont à l'origine de l'erreur E, pourtant selon r_1, l'erreur E est la seule origine de l'erreur désignée par B. Donc, une attention particulière sur E seulement est nécessaire pour corriger la conception de l'élève par rapport à l'erreur B. On effectue le même raisonnement pour les deux règles approximatives $r_3 : AB \rightarrow E$ et $r_4 : B \rightarrow E$ de la réduction transitive de la base informative (3.6). Il faut aussi souligner que pour r_3 et r_4, nous avons :

$$\begin{aligned} \text{Supp}\,(r_4 : B \rightarrow E) &> \text{Supp}\,(r_3 : AB \rightarrow E)\,, \\ \text{Conf}\,(r_4 : B \rightarrow E) &> \text{Conf}\,(r_3 : AB \rightarrow E)\,. \end{aligned}$$

Il est clair que r_4 est beaucoup plus informative que r_3. Non seulement, elle a les plus grandes mesures, mais elle a aussi une prémisse minimale. Bref, entre les règles $r_1 : E \rightarrow B$, $r_2 : AE \rightarrow B$, $r_3 : AB \rightarrow E$ et $r_4 : B \rightarrow E$, les plus importantes sont : $r_1 : E \rightarrow B$ et $r_4 : B \rightarrow E$. C'est ainsi que Gasmi *et al.* introduit une nouvelle base générique des règles définie ci-dessous.

$$NBG = \{r : X \rightarrow Y \setminus X \text{ telle que } Y \in \mathcal{F}, X \in \mathcal{G}_Z, Z \in \mathcal{F}, Z \subseteq Y, \text{Conf}(r) \geqslant minConf, \nexists G \subset X \text{et Conf}(G \rightarrow Y \setminus G) \geqslant minConf\}$$

Précisons que la base NBG regroupe à la fois les règles exactes et les règles approximatives. En effet, dans une règle de type $X \rightarrow Y \setminus X$ avec $X \in \mathcal{G}_Z$ et $Z \subseteq Y$, lorsque $Z = Y$, alors on obtient une règle exacte. Cette base est un sous ensemble de la réunion de la base informative (BI) et la base générique des règle exacte (BG) en supprimant les règle dont la prémisse n'est pas minimale. C'est à dire, lorsque $X \rightarrow Y \setminus X$ est dans la NBG, alors on ne doit plus y trouver une règle de type $Z \rightarrow Y \setminus Z$ telle que $Z \supset X$. Nous allons voir les éléments de la nouvelle base générique du contexte binaire $\mathcal{K}$ du tableau 2.1. Pour un $minSupp = 1/2$, nous n'avons que trois règles qui ont des prémisses composées de plus de deux items (k-itemsets avec $k \geqslant 2$) : $AE \rightarrow B$, $AB \rightarrow E$ et $AB \rightarrow C$. Examinons une à une ces trois règles pour voir si elles font partie de NBG. En prenant le générateur AE, on peut obtenir la règle exacte $AE \rightarrow ABE \setminus AE$, c'est-à-dire la règle $AE \rightarrow B$. Or, à partir du générateur E ($E \subset AE$), on peut avoir la règle $E \rightarrow ABE \setminus E$ ($E \rightarrow AB$). Donc la règle $AE \rightarrow ABE \setminus AE$ n'appartient pas à NBG. Dans la base informative, nous avons les deux règles $AB \rightarrow ABE \setminus AB$ et $B \rightarrow ABE \setminus B$. Comme $B \subset AB$, la règle $AB \rightarrow ABE \setminus AB$ ne fait pas partie de NBG. Avec le même raisonnement, la présence des règles $AB \rightarrow ABC \setminus AB$ et $B \rightarrow ABC \setminus B$ dans la base informative permet d'affirmer que $AB \rightarrow C$ ($AB \rightarrow ABC \setminus AB$) n'appartient pas à NBG. D'où la liste des éléments de NBG du contexte binaire $\mathcal{K}$ du tableau 2.1.

$$NBG = \{A \rightarrow B; A \rightarrow BC; A \rightarrow D; A \rightarrow BE; B \rightarrow E; B \rightarrow AE; B \rightarrow A; B \rightarrow AC; B \rightarrow D; D \rightarrow B; D \rightarrow A; E \rightarrow AB\}$$

3.3.2 Bases des règles M_{GK} valides

Un des objectifs de nos travaux est de redéfinir les bases des règles M_{GK}-valides pour avoir non seulement un ensemble minimal des règles à partir duquel il est possible, via les axiomes d'inférence, de retrouver la totalité des règles valides, mais aussi de faire en sorte qu'elles soient constituées uniquement des règles les plus informatives. Les définitions des bases des règles valides selon la mesure M_{GK}, les axiomes d'inférence permettant de retrouver l'ensemble des règles ainsi que les algorithmes d'extraction des bases ont été proposés dans (Feno, 2007). Cet auteur a bien souligné que ces algorithmes n'étaient pas encore optimisés. Avant de pouvoir apporter une éventuelle amélioration à ces bases des règles, nous allons d'abord identifier les limites de ces algorithmes en analysant le fonctionnement de ces derniers et la qualité des règles extraites. Nous allons donc voir dans ce paragraphe un bref rappel de description des bases des règles valides selon la mesure M_{GK} définies dans (Feno *et al.*, 2006). Ensuite, nous analyserons la forme et l'informativité des règles constituant ces bases et, éventuellement, les algorithmes permettant de les extraire.

Règles positives exactes

Pour tout contexte binaire $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ et pour tous motifs X, Y de $\mathscr{P}(\mathcal{I})$, la mesure selon M_{GK} de la règle $r : X \rightarrow Y$ est toujours comprise entre -1 et 1 (M_{GK} est une mesure normalisée selon la définition 2.4). Une règle $r : X \rightarrow Y$ est appelée règle exacte valide selon la mesure M_{GK} lorsque X et Y sont des motifs fréquents ($\mathrm{Supp}(X) > minSupp$, $\mathrm{Supp}(Y) > minSupp$) et $\mathrm{M}_{GK}(X \rightarrow Y) = 1$. Selon la propriété 4.3, on a une équivalence entre les règles exactes M_{GK}-valides et les règles exactes valides selon la mesure Confiance ($\mathrm{M}_{GK}(X \rightarrow Y) = 1 \Leftrightarrow \mathrm{Conf}(X \rightarrow Y) = 1$). Cette propriété explique pourquoi on a pris la base de Duquenne-Guigues (voir § 3.3.1) comme base des règles exactes M_{GK}- valides.

$$BPE = \{r : X \rightarrow \gamma(X) \setminus X \text{ telle que } X \in PFF\}, \tag{3.12}$$

où PFF désigne l'ensemble des pseudo-fermés fréquents.
Donc, si on considère le contexte binaire $\mathcal{K}$ du tableau 2.1 et en prenant un $minSupp = 1/3$, la base des règles exactes M_{GK}-valides est formée par les règles dans le tableau 3.2. Or, selon la précision que nous avons donnée au paragraphe 3.3.1, la base de Duquenne-Guigues n'est pas constituée des règles plus informatives. Dans le cas où l'utilisateur se contente d'interpréter les règles dans la base (sans dériver les autres règles), il serait beaucoup plus intéressant de ne lui proposer que les règles les plus informatives.

Règles négatives exactes

Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte d'extraction. Pour tous motifs X, Y de $\mathscr{P}(\mathcal{I})$, lorsque X favorise $\overline{Y}$ (X défavorise Y), alors $\mathrm{M}_{GK}(X \rightarrow \overline{Y}) \in]0, 1]$. On dit que la règle $X \rightarrow \overline{Y}$ est une règle négative exacte lorsque $\mathrm{M}_{GK}(X \rightarrow \overline{Y}) = 1$. Avant de donner les caractérisations de la base des règles négatives exactes, nous allons d'abord voir la notion de bordure positive.

Définition 3.9. *Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte binaire d'extraction. On appelle bordure positive, l'ensemble noté par $bd^+(0)$ de motifs défini par :*

$$bd^+(0) = \{X \subseteq I,\ Supp(X) > 0 \text{ et } \forall\, x,\ x \notin X,\ Supp(X \cup \{x\}) = 0\}.$$

Propriété 3.10. *Soient X et Y deux motifs de supports non nuls. Nous avons une équivalence entre les deux égalités ci-dessous :*

$$\begin{aligned} \mathrm{M}_{GK}(X \rightarrow \overline{Y}) &= 1; \\ Supp(X \rightarrow Y) &= 0. \end{aligned}$$

Preuve.

$$\begin{aligned} M_{GK}\left(X \rightarrow \overline{Y}\right) = 1 &\Leftrightarrow \frac{P\left(\overline{Y'}/X'\right) - P\left(\overline{Y'}\right)}{1 - P\left(\overline{Y'}\right)} = 1 \\ &\Leftrightarrow P\left(\overline{Y'}/X'\right) - P\left(\overline{Y'}\right) = 1 - P\left(\overline{Y'}\right) \\ &\Leftrightarrow P\left(\overline{Y'}/X'\right) = 1 \\ &\Leftrightarrow P\left(Y'/X'\right) = 0 \\ &\Leftrightarrow \frac{\mathrm{Supp}\left(X \rightarrow Y\right)}{\mathrm{Supp}(X)} = 0 \\ &\Leftrightarrow \mathrm{Supp}\left(X \rightarrow Y\right) = 0 \end{aligned}$$

□

Cette équivalence se trouve à la base des deux axiomes d'inférence permettant de définir la base des règles négatives exactes.

Axiomes d'inférence
Pour tous motifs X, Y et Z d'un contexte binaire d'extraction $\mathcal{K}$, on a :

$$(RNE1) \quad Si \begin{cases} M_{GK}\left(X \rightarrow \overline{Y}\right) = 1 \\ \text{Supp}\left(Y \cup Z\right) > 0, \end{cases} \quad \text{alors } M_{GK}\left(X \rightarrow \overline{Y \cup Z}\right) = 1.$$

$$(RNE2) \quad Si \begin{cases} M_{GK}\left(X \rightarrow \overline{Y}\right) = 1 \\ Z \subset X \\ \text{Supp}\left(Z \cup Y\right) = 0, \end{cases} \quad \text{alors } M_{GK}\left(Z \rightarrow \overline{Y}\right) = 1.$$

À partir de ces axiomes, on peut définir la base des règles négatives exactes. Feno *et al.* ont montré que l'ensemble BNE défini ci-dessous est une base pour les règles négatives exactes M_{GK}-valides relativement aux axiomes d'inférence $RNE1$ et $RNE2$.

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

Dans (Totohasina, 2008), on a souligné que la bordure positive $Bd^{+}(0)$ est identique à l'ensemble des motifs fermés maximaux de Support strictement positif et la base BNE pour les règles négatives exactes peut être exprimée en utilisant l'opérateur de fermeture. À partir de cette remarque, on peut affirmer que si un motif X est dans $Bd^{+}(0)$, alors il est forcément un fermé. Donc, les prémisses des règles dans BNE sont des fermées, donc motifs de taille maximales par rapport aux tailles des motifs dans sa classe. Si on tient compte du fait qu'une règle est plus informative que d'autres lorsqu'elle a une prémisse minimale, il va de soi que la base BNE, telle qu'elle est définie actuellement, n'est pas constituée des règles plus informatives. C'est à dire, il peut y avoir des règles dérivées d'une règle r de BNE qui soient plus informatives que r. Prenons par exemple le contexte binaire $\mathcal{K}$ décrit dans le tableau 2.1. Les éléments de $Bd^{+}(0)$ sont : $ABCD; ABCEF; ABDEF$. Donc l'ensemble des règles négatives exactes est :

$$BNE = \left\{ABCD \rightarrow \overline{E}\,;\ ABCD \rightarrow \overline{F}\,;\ ABCEF \rightarrow \overline{D}\,;\ ABDEF \rightarrow \overline{C}\,\right\}.$$

À présent, examinons la classe [4] des éléments constituant la bordure positive. Nous pouvons

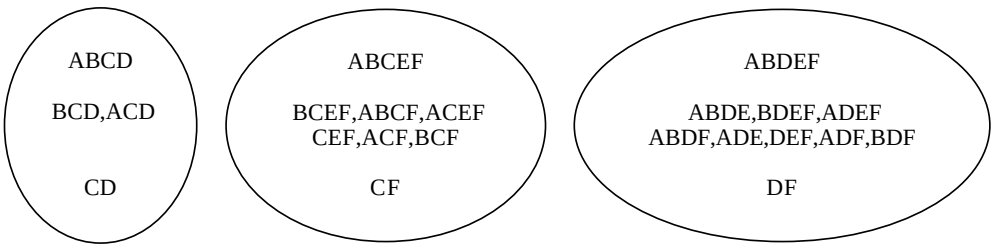


FIGURE 3.4 – Classe des éléments constituant Bd^{+}

voir dans la figure 3.4, que pour un X dans $Bd^{+}(0)$, la classe de X est formée par des

4. Ensemble des motifs de même fermé

éléments de taille plus petit que X. Or, on peut monter que pour tout X dans $Bd^+(0)$, si $X \to \overline{x}$ est dans BNE, alors pour tout Z dans la classe de X, $\mathrm{M}_{GK}(Z \to \{\overline{x}\}) = 1$, $Z \to \{\overline{x}\}$ est aussi une règle négative exacte. Dans l'exemple du contexte décrit dans le tableau 2.1, la règle $r_1 : ABCEF \to \overline{D}$ est dans BNE, pourtant, connaissant les éléments constituant la classe du motif fermé $ABCEF$, en l'occurrence son générateur CF, on peut affirmer que $r_2 : CF \to \overline{D}$ est aussi une règle négative exacte et de plus, on peut déduire toutes les règles dérivées de r_1 à partir de la règle r_2 ; et r_2 est visiblement plus informative que r_1. Ce constat nous amène à proposer une nouvelle description d'une base des règles négatives exactes.

Règles approximatives

Ce sont des règles dont les mesures M_{GK} varient dans l'ouvert $]-1, 0[\cup]0, 1[$. Les propriétés de motifs fermés ont permis de dégager des axiomes d'inférence pour les règles approximatives. En effet, pour tous motifs X, Y, Z, T tels que $Z \in [X]$ et $T \in [Y]$, il est facile de montrer que :

$$\begin{cases} M_{GK}(X \to Y) = M_{GK}(Z \to T), \\ M_{GK}\left(X \to \overline{Y}\right) = M_{GK}\left(Z \to \overline{T}\right). \end{cases}$$

À partir de ce constat, la validité des axiomes d'inférence (PA) et (NA) ci-dessous, axiomes permettant de retrouver les règles positives approximatives et les règles négatives approximatives est justifiée dans (Feno *et al.*, 2006 ; Feno, 2007).

(PA) Si $X \to Y$, Z et T étant deux motifs tels que $\varphi(Z) = \varphi(X)$ et $\varphi(T) = \varphi(Y)$, alors $Z \to T$.

(NA) Si $X \to \overline{Y}$, Z et T étant deux motifs tels que $\varphi(Z) = \varphi(X)$ et $\varphi(T) = \varphi(Y)$, alors $Z \to \overline{T}$.

À leur tour, ces axiomes d'inférence ont permis de définir la base positive approximative (BPA) et la base négative approximative (BNA) ci-dessous :

$$\begin{aligned} BPA(\alpha) &= \{X \to Y : \varphi(X) = X, \varphi(Y) = Y, Supp(Y)(1-\alpha) + \alpha \leq Conf(X \to Y) < 1\}, \\ BNA(\alpha) &= \{X \to \overline{Y} : \varphi(X) = X, \varphi(Y) = Y, 0 < Conf(X \to Y) \leq Supp(Y)(1-\alpha)\}. \end{aligned}$$

Il faut noter que par rapport aux axiomes d'inférence (PA) et (NA), les deux bases BPA(α) et BNA(α) sont minimales. Autrement dit, d'une part, on peut extraire toutes les règles valides à partir des éléments de la base et d'autre part, on ne peut plus réduire ces éléments sans perdre des règles valides au seuil α.

Algorithme d'extraction des bases BPA(α) et BNA(α). Dans le but d'apporter des améliorations aux qualités des règles constituant les bases des règles M_{GK}-valides, nous allons analyser le fonctionnement et les résultats fournis par l'algorithme d'extraction de $BPA(\alpha)$ et $BNA(\alpha)$ décrit dans (Feno, 2007). Plus précisément, nous allons apporter quelques critiques à l'algorithme 1. Ces critiques vont prendre deux formes : d'une part, au niveau purement algorithmique (forme de l'algorithme) et d'autre part, au niveau sémantique (forme des règles qui sortent de cet algorithme). Dans ce paragraphe, nous allons d'abord faire une description de cet algorithme ; ensuite, nous allons le dérouler pour essayer de voir ce qui en sort. Ces analyses seront faites dans le but de montrer que plusieurs points concernant les bases des règles approximatives (positives et négatives) peuvent encore être améliorés. Les notations utilisées dans l'algorithme 1 sont données dans le tableau 3.7. Pour fonctionner,

cet algorithme a besoin de la liste des fermés fréquents du contexte étudié. Cela suppose que les fermés fréquents ont été générés par d'autres algorithmes.

$NComp_i(X)$	Les motifs non comparables à X de taille inférieure ou égal à $\|X\|$
$SEns(FCj, X)$	Les fermés à la fois sous-ensembles de X et de FCj
FCi	Les i-motifs fermés fréquents
k	Taille maximale des motifs fermés fréquents

Tableau 3.7 – Notation et sous-programme

Algorithme 1 Bases des règles négatives et positives approximatives

Entrée : FCi Ensemble des fermés fréquents et leurs supports ; α : *seuil* M_{GK}
Sortie : BNA, BPA : Bases pour les règles négatives et positives approximatives

1: $BNA \leftarrow \{\}; BPA \leftarrow \{\}$
2: **Pour** $(FCi; i \leq k)$ **faire**
3: **Pour** $(Y\ dans\ FCi)$ **faire**
4: **Pour** $(X\ dans\ SEns(FCj, Y))$ **faire**
5: $Conf \longleftarrow Supp(X \cup Y)/Supp(X)$
6: **Si** $Supp(Y)(1-\alpha)+\alpha \leq Conf$ **alors**
7: $BPA \leftarrow BPA \cup \{X \rightarrow Y\}$
8: **Fin Si**
9: **Fin Pour**
10: **Pour** $(X\ dans\ NComp_i(Y))$ **faire**
11: $Conf \longleftarrow Supp(X \cup Y)/Supp(X)$
12: **Si** $0 < Conf \leq Supp(Y)(1-\alpha)$ **alors**
13: $BNA \longleftarrow BNA \cup \{X \rightarrow \overline{Y}, Y \rightarrow \overline{X}\}$
14: **Sinon**
15: **Si** $(Supp(Y)(1-\alpha)+\alpha \leq Conf < 1)$ **alors**
16: $BPA \longleftarrow BPA \cup \{X \rightarrow Y\}$
17: **Fin Si**
18: **Fin Si**
19: **Fin Pour**
20: **Fin Pour**
21: **Fin Pour**
22: **Retourner** BNA, BPA

Exécution de l'algorithme 1

Afin de découvrir le type des règles qui pourraient se trouver dans la base BPA-BNA, nous allons dérouler manuellement l'algorithme 1 en utilisant le contexte binaire d'extraction $\mathcal{K}$ (tableau 2.1). Remarquons que l'algorithme 1 exige que les fermés soient déterminés au préalable. En utilisant le « package arules » du logiciel R, nous sommes parvenus à trouver tous les fermés du contexte $\mathcal{K}$ (Tableau 2.1).

En prenant le seuil $\alpha = 0.1$, les fermés fréquents sont :

$$\begin{aligned} FC1 &= \{A, B, D\}, \\ FC2 &= \{AB, AD, BD, BE\}, \\ FC4 &= \{ABCD, ABCE, ABEF\}, \\ FC3 &= \{ABC, ABE, BDE, ABD\}, \\ FC5 &= \{ABDEF, ABCEF\}. \end{aligned}$$

Ici, la taille maximale d'un motif fréquent est : $k = 5$. Remarquons d'abord que pour deux motifs comparables X et Y ($X \subseteq Y$ ou $Y \subseteq X$), on a toujours une attraction mutuelle entre les deux motifs. En effet, supposons que $X \subseteq Y$, on a toujours $X \cup Y = Y$ et, par conséquent Supp($X \cup Y$) est égal à Supp(Y). En divisant Supp($X \cup Y$) par Supp(X) et Supp(Y) par n (nombre total des transactions) et en faisant remarquer que Supp(X)$\leq n$, on arrive à conclure que $P(Y' \setminus X') \geq P(Y')$. Donc, on a une attraction mutuelle entre les deux motifs, et par la suite, ce n'est pas avec des motifs (prémisse et conséquent) comparables qu'on retrouvera les règles négatives. Ceci a pour conséquence de restreindre le champ de recherche des règles négatives de la base sur les motifs fermés fréquents non comparables. D'où la necéssité de la fonction $NComp_i(Y)$ qui va retourner les motifs non comparables à Y et de taille inférieure ou égale à i. Pour notre contexte binaire $\mathcal{K}$, les motifs non comparables à un motif fermé sont donnés ci-après.

$$\begin{aligned} NComp_i(A) &= \{B, D\} \\ NComp_i(B) &= \{A, D\} \\ NComp_i(D) &= \{A, B\} \\ NComp_i(AB) &= \{D, AD, BD, BE\} \\ NComp_i(AD) &= \{B, AB, BD, BE\} \\ NComp_i(BD) &= \{A, AB, AD, BE\} \\ NComp_i(BE) &= \{A, D, AB, AD, BD\} \\ NComp_i(ABC) &= \{D, AD, BD, BE, ABE, BDE, ABD\} \\ NComp_i(ABE) &= \{D, AD, BE, ABC\} \\ NComp_i(ABCD) &= \{ABE\} \\ NComp_i(ABCE) &= \{D, AD, BD, ABCD, ABEF\} \\ NComp_i(ABEF) &= \{D, AD, BD, ABC, ABCD, ABCE\} \\ NComp_i(ABDEF) &= \{ABC, ABCD, ABCE, ABCEF\} \\ NComp_i(ABCEF) &= \{D, ABCD, ABDEF\} \end{aligned}$$

Faisons maintenant quelques pas dans l'exécution de l'algorithme afin de découvrir les points qui nécessitent des améliorations.

Pour $i = 1$ $FC1 = \{A, B, D\}$

Pour $Y = \{B\}$
$SEns(FC1, B) = \emptyset$
$NComp_i(B) = \{A, D\}$
$X = \{D\}$
$P(Y'/X') = 0,75$
$\text{Supp}(Y)(1-\alpha) = 0,75$
$\text{Supp}(Y)(1-\alpha) + \alpha = 0,85$
$BNA = \{D \to \overline{B}, B \to \overline{D}\}$

Pour $Y = \{D\}$
$SEns(FC1, D) = \emptyset$
$NComp_i(D) = \{A, B\}$
$X = \{B\}$
$P(Y'/X') = 0,6$
$\text{Supp}(Y)(1-\alpha) = 0,6$
$\text{Supp}(Y)(1-\alpha) + \alpha = 0,7$
$BNA = \{B \to \overline{D}, D \to \overline{B}\}$

Pour $i = 2$ $FC2 = \{AB, AD, BD, BE\}$

Pour $Y = \{AD\}$
$SEns(FC2, AD) = \{A, D\}$
$NComp_i(AD) = \{B, AB, BD, BE\}$
$X = \{BE\}$
$P(Y'/X') = 0,25$
$\text{Supp}(Y)(1-\alpha) = 0,45$
$\text{Supp}(Y)(1-\alpha) + \alpha = 0,55$
$BNA = \{BE \to \overline{AD}, AD \to \overline{BE}\}$

Pour $Y = \{BE\}$
$SEns(FC2, BE) = \{B\}$
$NComp_i(BE) = \{A, D, AD, AB, BD\}$
$X = \{AD\}$
$P(Y'/X') = 0,33$
$\text{Supp}(Y)(1-\alpha) = 0,6$
$\text{Supp}(Y)(1-\alpha) + \alpha = 0,7$
$BNA = \{AD \to \overline{BE}, BE \to \overline{AD}\}$

Pour $i = 3$: $FC3 = \{ABC, \ldots, BDE\}$
Pour $Y = \{ABC\}$
$SEns(FC3, ABC) = \{A, B, AB\}$
$NComp_i(ABC) = \{D, AD, \ldots, ABD\}$
$X = \{B\}$
$P(Y'/X') = 0,6$
$\text{Supp}(Y)(1-\alpha) = 0,45$
$\text{Supp}(Y)(1-\alpha) + \alpha = 0,55$
$BPA = \{B \to ABC\}$

Pour $i = 4$: $FC4 = \{ABCD, ABCE, ABEF\}$
Pour $Y = \{ABCE\}$
$SEns(FC4, ABCE) = \{A, B, BE, \ldots, ABE\}$
$NComp_i(ABCE) = \{D, AD, \ldots, ABEF\}$
$X = \{B\}$
$P(Y'/X') = 0,4$
$\text{Supp}(Y)(1-\alpha) = 0,3$
$\text{Supp}(Y)(1-\alpha) + \alpha = 0,4$
$BPA = \{B \to ABCE\}$

Critiques sur les bases des règles approximatives M_{GK} valides
Ces quelques pas d'exécution dans l'algorithme 1 illustrent les résultats des analyses critiques à l'égard de la description des bases des règles approximatives que nous allons effectuer. Ces résultats peuvent être classés en deux catégories : les critiques relatives à la conception de l'algorithme et ceux qui sont liées à la forme, donc à la qualité des règles constituant la base. Nous allons détailler l'un après l'autre, les points qui nous semblent importants à améliorer.

Premier cas : Nous savons que l'un des objectifs de l'extraction des bases des règles est de trouver un ensemble minimal (en terme de cardinal et relativement à un ensemble d'axiome d'inférence) à partir duquel, via les axiomes d'inférences, on peut, si l'on veut, retrouver la totalité des règles valides. Nous avons montré que dans le cas d'attraction mutuelle entre le conséquent et la prémisse, la mesure M_{GK} vérifie la propriété sur la contraposition de la logique classique. C'est à dire, si X favorise Y, alors $\text{M}_{GK}(X \to Y) = \text{M}_{GK}(\overline{Y} \to \overline{X})$, une

égalité qui est tout à fait cohérente avec l'équivalence entre une proposition et sa contraposée ($p \Rightarrow q \Leftrightarrow \overline{q} \Rightarrow \overline{p}$). Suite à ces constats, mettre $Y \rightarrow \overline{X}$ dans la base alors qu'on a déjà, dans cette même base, la règle $X \rightarrow \overline{Y}$ ne fait que doubler inutilement le nombre des règles constituant la base des règles négatives approximatives. En effet, si $X \rightarrow \overline{Y}$ est dans $BNA(\alpha)$, alors on a nécessairement une attraction mutuelle entre X et $\overline{Y}$ (X favorise $\overline{Y}$ qui est d'ailleurs équivalent à X défavorise Y). Dans ce cas, en appliquant la propriété d'équivalence d'une proposition avec sa contraposée, à partir de la règle $X \rightarrow \overline{Y}$, on peut facilement déduire la règle $Y \rightarrow \overline{X}$.
Toujours au niveau de la forme de l'algorithme 1, nous pouvons remarquer qu'aux différents niveaux de l'algorithme, on peut obtenir exactement les mêmes règles. La suppression de ce type de répétition pourrait apporter une nette amélioration aux temps d'exécution de l'algorithme.
Enfin, nous avons pu remarquer que dans cet algorithme, la validation d'une règle est effectuée en comparant la valeur M_{GK} de la règle en question au seuil α fixé par l'utilisateur. Rappelons que ce même seuil est utilisé dans la sélection des motifs fréquents. Or, selon (Feno *et al.*, 2006), la validation d'une règle selon M_{GK} doit se faire en comparant la mesure de la règle avec la valeur critique de M_{GK} calculée au seuil fixé par l'utilisateur. Il est donc nécessaire d'insérer le calcul des valeurs critiques dans cet algorithme afin que l'on puisse l'utiliser pour valider ou rejeter des règles. Quelques modifications dans le pseudo-code permettant d'extraire les bases des règles approximatives M_{GK}-valides pourraient améliorer les points que nous venons de citer. Nous allons maintenant voir la forme et analyser la qualité des règles extraites par l'algorithme 1.

Deuxième cas : Puisque la prémisse et le conséquent sont à choisir dans l'ensemble des motifs fermés fréquents, ceux-ci ne sont pas nécessairement disjoints. Au contraire, dans la majorité des cas, ils ont une intersection non vide. Donc, si on met la règle $r : X \rightarrow Y$ (avec X, Y fermés) dans la base, on risque de contredire la définition selon laquelle une règle d'association est une implication entre deux motifs disjoints. De plus, il serait difficile d'interpréter une règle négative à prémisse et conséquent non disjoints. Pour avoir une idée de ce que cela peut être, il suffit d'imaginer l'interprétation de la règle $XY \rightarrow \overline{YZ}$. Soulignons que contrairement aux mesures Support et Confiance, avec la mesure M_{GK}, les deux règles $XY \rightarrow \overline{YZ}$ et $XY \rightarrow \overline{Z}$ ne s'interprètent pas de la même manière pour la simple raison qu'elle n'ont pas forcement la même mesure (cf. Fig 4.2 du § 4.6.1 sur l'étude des variations de M_{GK} par rapport au support de conséquent), ni la même valeur critique.
Par rapport au concept de redondance des règles, plusieurs textes présents dans la littérature, en l'occurrence (Pasquier, 2000a ; Bastide *et al.*, 2002 ; Gasmi *et al.*, 2006), affirment qu'entre deux règles de même intensité (c'est-à-dire de mêmes mesures) et de prémisse et conséquent comparables, la plus informative est celle qui a une prémisse minimale et un conséquent maximal. Comme les fermés sont souvent comparables, il est très fréquent de rencontrer dans une base des règles approximatives, des règles de types $X \rightarrow Y$ et $Z \rightarrow T$, avec $X \subseteq Z$ ou $Y \supseteq T$. C'est par exemple le cas pour les règles $r_1 : B \rightarrow ABC$ et $r_2 : B \rightarrow ABCE$ que nous venons d'avoir lors de l'exécution de l'algorithme 1. Certes, les règles r_1 et r_2 pourraient avoir des mesures différentes, mais si on se met à la place de l'utilisateur, qui est prêt à prendre en compte au même pied les informations valides au niveau de confiance α qu'il vient de fixer, il serait légitime de se demander l'utilité de la règle r_2 lorsqu'on a déjà r_1. Nous revenons à cette question lors de la définition de semi-base des règles M_{GK}-valides.

Nous pouvons aussi remarquer que l'algorithme 1 prend comme données en entrée l'ensemble des motifs fermés fréquents. C'est-à-dire, pour un *minSupp* fixé, tout motif X dans FCi vérifie l'inégalité $\text{Supp}(X) \geq minSupp$. Cet algorithme fournit en sortie la base positive et négative approximative (BPA, BNA). La base BNA est constituée par des règles de type $X \rightarrow \overline{Y}$, où X et Y sont des motifs fermés fréquents. Pourtant, si $\text{Supp}(Y)$ est plus grand que le *minSupp*, il n'en est pas toujours le cas pour le support de $\overline{Y}$ ($\text{Supp}(\overline{Y}) = 1 - \text{Supp}(Y)$). Autrement dit, dans sa forme actuelle, il est tout à fait possible de trouver dans BNA des règles constitués des motifs non fréquents. D'un autre côté, puisqu'à l'entrée de l'algorithme, on a seulement pris les fermés fréquents, on peut donc perdre des règles négatives constituées par des motifs de type $\overline{X}$ dont les motifs positifs X ne sont pas fréquents. Pour illustrer ces propos, nous allons considérer le contexte binaire d'extraction donné dans le tableau 3.8 ci-dessous. Tout d'abord, soulignons que les motifs X, Y, Z et T sont tous des motifs fermés.

X	Y	Z	T
1	0	1	0
1	0	1	0
1	0	1	0
1	1	1	0
0	1	1	0
0	1	1	0
1	1	0	0
0	1	1	1
1	0	0	1
0	1	0	1

Tableau 3.8 – Génération des règles négatives à partir des motifs positifs fréquents

En prenant un $minSupp = 0,5$, on peut voir facilement que X, Y, Z sont des motifs fréquents et, T ne l'est pas, puisque :

$$\begin{cases} Supp\,(X) = 0,6 \\ Supp\,(Y) = 0,6 \end{cases} \quad \text{et} \quad \begin{cases} Supp\,(Z) = 0,7 \\ Supp\,(T) = 0,3. \end{cases}$$

Examinons la règle $X \rightarrow Y$

$$\begin{aligned} P(Y'/X') - P(Y') &= \frac{n_{XY}}{n_X} - \frac{n_Y}{n} \\ &= \frac{2}{6} - \frac{6}{10} \\ &< 0. \end{aligned}$$

Donc, le motif X défavorise Y. Selon l'algorithme 1, puisque X et Y sont des fermés fréquents et X défavorise Y, on doit examiner la validité de la règle $X \rightarrow \overline{Y}$.

$$\begin{aligned} M_{GK}\left(X \to \overline{Y}\right) &= \frac{\frac{n_{X\overline{Y}}}{n_X} - \frac{n_{\overline{Y}}}{n}}{1 - \frac{n_{\overline{Y}}}{n}} \\ &= \frac{\frac{4}{6} - \frac{4}{10}}{1 - \frac{4}{10}} \\ &= 0,44. \end{aligned}$$

Prenons la précision $\alpha = 0,9$ pour évaluer la valeur critique [5] de M_{GK}. Au niveau de confiance 90% (risque d'erreur 10%), la valeur théorique de χ^2 est égal à 2, 71. Connaissant cette valeur théorique, on peut évaluer la valeur critique de M_{GK} selon l'expression :

$$\begin{aligned} M^{\alpha}_{GK}\left(X \to \overline{Y}\right) &= \sqrt{\frac{1}{n}\frac{n - n_X}{n_X}\frac{n_{\overline{Y}}}{n - n_{\overline{Y}}}\chi^2_{\text{Théorique}}} \\ &= \sqrt{\frac{1}{10}\frac{10 - 6}{6}\frac{4}{10 - 4} \times 2,71} \\ &= 0,34. \end{aligned}$$

Selon l'algorithme 1, comme la valeur de M_{GK} dépasse sa valeur critique au niveau de confiance 90%, la règle $X \to \overline{Y}$ doit être ajoutée à $BNA(\alpha)$. Pourtant le motif $\overline{Y}$ n'est même pas un motif fréquent, car $Supp(\overline{Y}) = 1 - Supp(Y) = 1 - 0,6 = 0,4 < 0,5$.

Examinons la règle $Z \to \overline{T}$

Comme T n'est pas un fermé fréquent (T est fermé, mais $\mathrm{Supp}(T) < minSupp$), il ne fera pas partie de ceux qui seront proposées à l'entrée de l'algorithme 1. Donc, la validité de la règle $Z \to \overline{T}$ ne sera même pas étudiée. Cependant, si on l'examine de près, on a :

$$\begin{aligned} M_{GK}\left(Z \to \overline{T}\right) &= \frac{\frac{n_{Z\overline{T}}}{n_Z} - \frac{n_{\overline{T}}}{n}}{1 - \frac{n_{\overline{T}}}{n}} \\ &= \frac{\frac{6}{7} - \frac{7}{10}}{1 - \frac{7}{10}} \\ &= 0,523. \end{aligned}$$

Or la valeur critique de la règle $Z \to \overline{T}$ est :

$$\begin{aligned} M^{\alpha}_{GK}\left(Z \to \overline{T}\right) &= \sqrt{\frac{1}{n}\frac{n - n_Z}{n_Z}\frac{n_{\overline{T}}}{n - n_{\overline{T}}}\chi^2_{\text{Théorique}}} \\ &= \sqrt{\frac{1}{10}\frac{10 - 7}{7}\frac{7}{10 - 7} \times 2,71} \\ &= 0,520. \end{aligned}$$

De plus, Z et $\overline{T}$ sont des motifs fréquents ($\mathrm{Supp}(\overline{T}) = 1 - \mathrm{Supp}(T) = 0,7$). En prenant un $minSupp = 50\%$ et au niveau de confiance $\alpha = 90\%$, les deux motifs Z et $\overline{T}$ sont

5. La notion de valeur critique sera détaillée au chapitre 4.

fréquents et $\mathrm{M}_{GK}(Z \rightarrow \overline{T}) > \mathrm{M}^{Cr}_{GK}(Z \rightarrow \overline{T})$, donc la règle $Z \rightarrow \overline{T}$ est bel et bien une règle valide. On peut donc constater que le fait de choisir seulement les fermés fréquents comme arguments à l'entrée de l'algorithme d'extraction de $BPA(\alpha)$ et $BNA(\alpha)$ peut nous amener à valider une règle négative composée des motifs non fréquents comme il peut nous amener à perdre des règles négatives valides que l'on peut extraire des motifs positifs non fréquents et, évidemment avec elles, toutes leurs règles dérivées. Il est donc important de tenir compte de cette dualité entre le support d'un motif X et celui de $\overline{X}$ relativement à un *minsupp* fixé. Nous y reviendrons lors de la conception des algorithmes d'extraction des nouvelles bases des règles M_{GK}-valides.

3.4 Conclusion partielle

Ce chapitre nous a permis de voir quelques bases des règles qu'on peut retrouver dans la littérature. Les principales idées autour desquelles est définie chacune de ces bases sont : la réduction dans la mesure du possible du nombre des règles constituant les bases, la possibilité de retrouver l'ensemble des règles valides via des axiomes d'inférence adéquats et enfin, la quantité d'information véhiculée par les éléments de chaque base. Étant donnés deux motifs X et Y d'un certain contexte binaire d'extraction $\mathcal{K}$ et une règle $r : X \rightarrow Y$, les bases des règles peuvent être divisées en deux grandes catégories : bases des règles exactes (composées des règle de type $X \rightarrow Y$ lorsque $X' \subseteq Y'$, pour certaine mesure μ, on a $\mu(r : X \rightarrow Y) = 1$) et les bases des règles approximatives (composées des règle de type $X \rightarrow Y$ lorsque $X' \nsubseteq Y'$, pour certaine mesure μ, on a $\mu(r : X \rightarrow Y) < 1$). Pour pouvoir établir des axiomes d'inférence, ces bases doivent être définies à partir des motifs particuliers comme les pseudos-fermés, les fermés, les générateurs et les bordures positives. À travers des exemples, nous avons vu que la constitution des bases des règles permet de synthétiser les règles valides, donc de réduire considérablement les règles à interpréter et ceci doit se faire sans aucune perte d'information utile.
Nous avons énuméré au chapitre 2 un certain nombre d'arguments qui nous ont poussé à choisir l'utilisation de la mesure de qualité M_{GK} dans l'extraction des règles d'association. Pour automatiser l'extraction et rendre possible l'interprétation des règles M_{GK}-valides, il est nécessaire de passer par l'extraction des bases des règles. Telles qu'elles sont définies actuellement et par rapport aux propriétés mathématiques (comme la disjonction entre la prémisse et conséquent d'une règle, la représentativité des motifs (positif et négatif) par rapport à un *minSupp* fixé au préalable...) et sémantiques (minimalité de prémisse, maximalité de conséquent d'une règle positive, minimalité de conséquent d'une règle négative...) des bases des règles existant dans la littérature, nous avons soulevé plusieurs points que l'on peut améliorer dans le but de disposer des règles M_{GK}-valides beaucoup plus fiables. Le prochain chapitre de ce rapport sera consacré exclusivement à la proposition des nouvelles descriptions des bases des règles M_{GK}-valides.

Deuxième partie

Contributions : ASI et didactique de mathématiques

Chapitre 4

Proposition des nouvelles bases des règles M_{GK}-valides

Sommaire

4.1 Introduction

Compte tenu des critiques avancées dans le précédent chapitre et dans l'objectif de toujours vouloir améliorer les éventuelles connaissances que l'on peut extraire d'une base des données, nous allons définir deux catégories de bases des règles M_{GK}-valides. Dans un premier temps, nous proposons une simple amélioration des bases définies dans (Feno *et al.*, 2006) en faisant en sorte que les prémisses et les conséquents soient toujours disjoints et de n'extraire que des règles plus informatives, c'est-à-dire des règles qui ont des prémisses minimales et conséquents

maximaux. Nous avons aussi utilisé la notion de valeur critique décrite dans (Totohasina *et al.*, 2004 ; Totohasina et Feno, 2008) pour valider ou non une règle. Précisons qu'à partir de ces nouvelles bases M_{GK}-valides, on peut toujours se servir des axiomes d'inférence décrits dans Feno *et al.* pour valider l'ensemble des règles valides. Il est bien vrai qu'un motif fermé est toujours maximal dans sa classe (rappelons qu'ici, une classe désigne l'ensemble des motifs ayant la même fermeture) et choisir un motif fermé comme conséquent nous donne une impression qu'on a toujours une règle à conséquent maximal. Cependant, si on examine l'ensemble des motifs fermés, ils sont souvent comparables ; il n'est donc pas étonnant que dans une base à conséquent fermé, on trouve encore des règles dont le conséquent n'est pas maximal. C'est ainsi qu'une réduction des bases des règles M_{GK}-valides sera proposée. Cette réduction est basée sur un raisonnement similaire à celui qui est utilisé par Gasmi *et al.* quand ils ont supprimé dans la base générique des règles exactes (BG) et dans la base informative des règles approximatives (BI) les règles dont les prémisses ne sont pas minimales. Dans la réduction d'une base des règles M_{GK}-valides que nous avons appelé *semi-base des règles*, nous allons encore plus loin en supprimant les règles dont les conséquents ne sont pas maximaux (même s'ils sont fermés). Avant de voir ces deux catégories de base des règles, nous allons d'abord décrire comment choisir les prémisses, les conséquents et les valeurs critiques utilisées pour définir les bases M_{GK}-valides.

4.2 Prise en compte des valeurs critiques de M_{GK}

Cette section a pour objectif de montrer l'utilité du test d'indépendance de χ^2 dans la validation d'un lien implicatif entre deux motifs. Par la même occasion, nous allons justifier le choix des valeurs critiques de M_{GK}.
Nous avons pu établir dans le chapitre 3 (§ 2.8) la conséquence engendrée par la simple utilisation de la probabilité conditionnelle sans tenir compte de la distance à l'indépendance. La courbe représentant les valeurs de M_{GK} en fonction du support de conséquent (Fig. 4.2) montre que pour des règles de même confiance (par exemple), il est possible d'avoir des motifs dépendants, indépendants ou même des motifs qui se repoussent l'un et l'autre. Tout dépend de la distance à l'indépendance (i. e. $P(Y'/X') - P(Y')$). Autrement dit, pour une « confiance » fixée, le lien entre deux motifs X et Y dépend de la valeur du support du conséquent ($\mathrm{Supp}(Y)$). Comme la quantité $P(Y'/X') - P(Y')$ décroit en fonction de support de Y, il est légitime de se demander « jusqu'à quelle valeur de $\mathrm{Supp}(Y) = P(Y')$ on peut accepter la dépendance entre deux motifs X et Y ».
Il est donc important dans le processus d'extraction des règles d'association de pouvoir établir que la dépendance observée est statistiquement significative. Tout ceci constitue des raisons pour lesquelles le test d'indépendance de χ^2 doit utilement entrer dans la validation d'une règle d'association. Il est utilisé pour valider si oui ou non, à un certain seuil fixé, le lien entre les motifs est ainsi significatif. En effet, il est logiquement préférable que pour n'importe quel seuil de confiance, le résultat fourni par une quelconque mesure vis-à-vis d'un éventuel lien entre deux motifs soit cohérent avec le test d'indépendance de χ^2. Pour mettre en évidence la nécessité d'un test d'indépendance par rapport aux mesures « confiance » et M_{GK}, nous allons considérer les deux situations fictives ci-après (cf. Tableau 4.1).
Après calcul, nous avons obtenu les valeurs ci-dessous.

	Y	$\overline{Y}$	$\sum$
X	1200	100	1300
$\overline{X}$	300	600	900
$\sum$	1500	700	2200

	T	$\overline{T}$	$\sum$
Z	1775	300	2075
$\overline{Z}$	100	25	125
$\sum$	1875	325	2200

Tableau 4.1 – Exemples comparatifs

Pour la règle $(X \to Y)$			**Pour la règle** $(Z \to T)$		
$\text{Conf}(X \to Y)$	=	0, 92	$\text{Conf}(Z \to T)$	=	0, 94
$\text{M}_{GK}(X \to Y)$	=	0, 75	$\text{M}_{GK}(Z \to T)$	=	0, 06
$\chi^2_{\text{Observé}}$	=	852	$\chi^2_{\text{Observé}}$	=	2, 87
Risque d'erreur (α)	=	5%	Risque d'erreur (α)	=	5%
$\chi^2_{\text{Théorique}}$	=	3, 84	$\chi^2_{\text{Théorique}}$	=	3, 84

Comme nous pouvons le constater, les règles $X \to Y$ et $Z \to T$ ont toutes les deux une valeur très élevée de la mesure confiance (0, 92 et 0, 94). Si on se contentait d'utiliser seulement cette mesure, on pourrait croire que les deux règles étaient intéressantes ; pire encore, on pourrait même être amené à croire que la règle $Z \to T$ est plus significative par rapport à la règle $X \to Y$ ($\text{Conf}(Z \to T) > \text{Conf}(X \to Y)$). Pourtant, le test d'indépendance de χ^2 montre clairement qu'au risque d'erreur de moins de 5%, on doit accepter l'hypothèse selon laquelle les deux motifs Z et T étaient indépendants, autrement dit, il n'existe aucune liaison entre Z et T ($\chi^2_{\text{Observé}} << \chi^2_{\text{Théorique}}$). Ce simple contre-exemple montre qu'on ne doit pas se contenter seulement de la mesure confiance pour valider une règle ; après tout, la dite mesure confiance n'est rien d'autre que la probabilité conditionnelle sachant la prémisse du motif conséquent.
D'un autre côté, utiliser seulement le test d'indépendance de χ^2 ne permet pas de distinguer la cause et la conséquence. En effet, le fait qu'il est symétrique ($\chi^2(X, Y) = \chi^2(Y, X)$), avec la valeur de χ^2, on peut juste affirmer que les deux motifs sont dépendants ou non, mais dans le cas de dépendance, on ne peut pas savoir la nature de cette dépendance (attractive ou bien répulsive, qui serait la cause ? et qui serait l'effet ?).
La question qui se pose à présent est, « doit-on toujours utiliser le test d'indépendance de χ^2 pour vérifier une règle déjà valide pour une quelconque mesure ? » Peut-être, le cas de la mesure confiance est assez clair, en soi, elle pourrait induire l'utilisateur en erreur parce qu'elle (la dite mesure confiance) peut valider des règles entre deux motifs statistiquement indépendants. Heureusement, cette question ne se pose plus pour le cas de la mesure M_{GK}. Voyons pourquoi.

Valeurs critiques de M_{GK}

En 2004, Totohasina *et al.* ont publié un lien direct entre la mesure M_{GK}, alors appelée **Ion** à l'époque (**I**mplication statistique **O**rientée **N**ormée) et l'expression de χ^2. Rappelons ici la proposition qui justifie cette relation (Totohasina *et al.*, 2004).

Propriété 4.1 (Relation entre χ^2 et M_{GK})**.** *Pour tous motifs X et Y, on a :*

$$M_{GK}(X \to Y) = \begin{cases} \sqrt{\frac{1}{n}\frac{n-n_x}{n_x}\frac{n_y}{n-n_y}\chi^2}, & \text{si } X \text{ favorise } Y \text{ ou } X \text{ et } Y \text{ indépendants,} \\ -\sqrt{\frac{1}{n}\frac{n-n_x}{n_x}\frac{n-n_y}{n_y}\chi^2}, & \text{si } X \text{ défavorise } Y \text{ ou } X \text{ et } Y \text{ indépendants.} \end{cases} \tag{4.1}$$

Où n, n_x, n_y désignent respectivement la taille des données, l'effectif des objets contenant le motif X et l'effectif des objets contenant le motif Y (cf. Tableau 4.2).

Preuve.
Considérons le tableau de contingence pour les deux motifs X et Y. Ici, n_{xy} désigne l'effectif des objets contenant à la fois le motif X et Y, pour les notations (voir § 3.2).

	Y	$\overline{Y}$	$\sum$
X	n_{xy}	$n_{x\overline{y}}$	n_x
$\overline{X}$	$n_{\overline{x}y}$	$n_{\overline{x}\,\overline{y}}$	$n_{\overline{x}}$
$\sum$	n_y	$n_{\overline{y}}$	n

Tableau 4.2 – Effectifs observés

Sous l'hypothèse d'indépendance entre X et Y (hypothèse nulle) on devrait avoir le tableau de contingence ci-dessous (cf. Tableau 4.3).

	Y	$\overline{Y}$	$\sum$
X	$\frac{n_x n_y}{n}$	$\frac{n_x n_{\overline{y}}}{n}$	n_x
$\overline{X}$	$\frac{n_{\overline{x}} n_y}{n}$	$\frac{n_{\overline{x}} n_{\overline{y}}}{n}$	$n_{\overline{x}}$
$\sum$	n_y	$n_{\overline{y}}$	n

Tableau 4.3 – Effectifs théoriques

Dans ce cas, la valeur théorique de χ^2 s'écrit :

$$\chi^2_{\text{Observé}} = \frac{\left(n_{xy} - \frac{n_x n_y}{n}\right)^2}{\frac{n_x n_y}{n}} + \frac{\left(n_{x\overline{y}} - \frac{n_x n_{\overline{y}}}{n}\right)^2}{\frac{n_x n_{\overline{y}}}{n}} + \frac{\left(n_{\overline{x}y} - \frac{n_{\overline{x}} n_y}{n}\right)^2}{\frac{n_{\overline{x}} n_y}{n}} + \frac{\left(n_{\overline{x}\overline{y}} - \frac{n_{\overline{x}} n_{\overline{y}}}{n}\right)^2}{\frac{n_{\overline{x}} n_{\overline{y}}}{n}}. \tag{4.2}$$

En se servant d'un outil de calcul formel, en l'occurrence le logiciel SAGE [1], on peut montrer facilement qu'en multipliant l'expression (4.2) par $\frac{1}{n}\frac{n-n_x}{n_x}\frac{n_y}{n-n_y}$, et en prenant la racine carrée, on tombe sur l'expression de $\mathrm{M}_{GK}(X \to Y)$ dans le cas de X favorisant Y. D'autre

1. http://www.sagemath.org/

part, si on multiplie l'expression (4.2) par $\frac{1}{n}\frac{n-n_x}{n_x}\frac{n-n_y}{n_y}$, et en prenant la racine carrée, on retrouve l'expression de $\mathrm{M}_{GK}(X \to Y)$ dans le cas de X défavorisant Y. □

Examinons maintenant le comportement de M_{GK} par rapport aux valeurs de χ^2. Au risque d'erreur α (erreur de première espèce[2]), on peut rejeter l'hypothèse nulle lorsque :

$$\chi^2_{\text{Observé}} > \chi^2_{\text{Théorique}}. \tag{4.3}$$

Or, rejeter l'hypothèse nulle signifie rejeter l'indépendance. À partir de l'inégalité (4.3), on peut obtenir :

$$\sqrt{\frac{1}{n}\frac{n-n_x}{n_x}\frac{n_y}{n-n_y}\chi^2_{\text{Observé}}} > \sqrt{\frac{1}{n}\frac{n-n_x}{n_x}\frac{n_y}{n-n_y}\chi^2_{\text{Théorique}}}.$$

Donc, d'après (4.1), on a (dans le cas de X favorisant Y) :

$$\mathrm{M}_{GK}(X \to Y) > \sqrt{\frac{1}{n}\frac{n-n_x}{n_x}\frac{n_y}{n-n_y}\chi^2_{\text{Théorique}}}. \tag{4.4}$$

Ici, $\chi^2_{\text{Théorique}}$ désigne la valeur théorique de χ^2 à 1 degré de liberté et à un risque d'erreur fixé par l'utilisateur. Remarquons qu'on peut se contenter de l'expression de M_{GK} dans le cas d'attraction mutuelle entre prémisse et conséquent. En effet, dans le cas de répulsion (X défavorisant Y par exemple), grâce à la propriété 3.6, on n'a qu'à s'intéresser à la règle $X \to \overline{Y}$) et utiliser l'expression de M_{GK} dans le cas d'attraction mutuelle. L'inégalité (4.4) nous amène aux définitions ci-dessous.

Définition 4.1 (**Valeur critique**). *Soient X et Y deux motifs, n_x et n_y deux entiers tels que : $n_x = Card(X'), n_y = Card(Y')$. On appelle valeur critique de M_{GK} au seuil α, la quantité notée par M^{α}_{GK} et définie par :* $\mathrm{M}^{\alpha}_{GK} = \sqrt{\frac{1}{n}\frac{n-n_x}{n_x}\frac{n_y}{n-n_y}\chi^2_{Théorique}(\alpha)}$.

Définition 4.2 (Règle valide). *Soit X et Y deux motifs fréquents*[3] *d'un contexte binaire d'extraction $\mathcal{K}$. On dit qu'une règle $X \to Y$ est valide au sens de la mesure M_{GK}, au risque d'erreur α lorsque $\mathrm{M}_{GK}(X \to Y) > M^{\alpha}_{GK}$.*

Remarque sur l'utilisation de la valeur critique

Il faut souligner que si on veut avoir un résultat fiable au niveau de confiance plus de $100\alpha\%$ (risque d'erreur moins de $100(1-\alpha)\%$) lors de la prise de décision concernant la validité d'une règle $X \to Y$, on ne doit pas comparer la valeur de M_{GK} à la quantité $100\alpha\%$ comme c'est le cas dans la plupart des algorithmes d'extraction des règles d'association, en l'occurrence l'ancien algorithme d'extraction des bases des règles M_{GK} valides (Feno, 2007) mais on doit plutôt la comparer à la quantité $M^{(1-\alpha)}_{GK}$. En effet, l'utilisation de $M^{(1-\alpha)}_{GK}$ permet de fournir à la mesure M_{GK} une valeur précise qu'il faut dépasser pour assurer l'existence de dépendance

2. Rejeter l'hypothèse nulle alors qu'elle est vraie.
3. Par rapport à un *minsup* préalablement établi.

significative entre X et Y selon le test statistique de χ^2.
Pour illustrer cette affirmation, prenons l'exemple de la règle $X \to Y$ d'un contexte dressé dans le tableau 4.1. Supposons que l'utilisateur souhaite avoir une règle valide avec une précision [4] plus grande que $\alpha = 0,95$. Si l'utilisateur compare $\mathrm{M}_{GK}(X \to Y)$ à la valeur $0,95$, il risque de perdre la règle $X \to Y$. Pourtant avec la même précision (risque d'erreur de première espèce moins de $(1-\alpha) = 0,05$), le test d'indépendance de χ^2 valide la liaison entre X et Y. Donc, l'utilisateur ne doit pas comparer la valeur de M_{GK} à la probabilité $\alpha = 0,95$ mais plutôt à la quantité $M_{GK}^{(1-\alpha)}$ (valeur critique de M_{GK} au risque d'erreur $(1-\alpha)$. Donc pour notre exemple, $\mathrm{M}_{GK}(X \to Y)$ est largement plus grand que $M_{GK}^{0,05}$ ($\mathrm{M}_{GK}(X \to Y) = 0,75 >> M_{GK}^{0,05} = 0,05088$. Donc, pour une précision de plus de $0,95$, la règle $X \to Y$ est valide au sens de la mesure M_{GK}. Nous allons donc utiliser cette expression de M_{GK}^{α} dans les nouveaux algorithmes d'extraction des règles M_{GK} valides.

Proposition 4.1. *Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte d'extraction. Pour tous motifs X, Y de $\mathscr{P}(\mathcal{I})$ et pour tout $\alpha \in]0,1[$, on a :*

1. *Si X favorise Y, alors $\mathrm{M}_{GK}^{\alpha}(X \to Y) = \mathrm{M}_{GK}^{\alpha}(\overline{Y} \to \overline{X})$,*
2. *Si X défavorise Y, alors $\mathrm{M}_{GK}^{\alpha}(X \to \overline{Y}) = \mathrm{M}_{GK}^{\alpha}(Y \to \overline{X})$.*

Preuve.
À un seuil d'erreur α, si X favorise Y, alors :

$$\begin{aligned} \mathrm{M}_{GK}^{\alpha}(X \to Y) &= \sqrt{\frac{1}{n}\frac{n-n_X}{n_X}\frac{n_Y}{n-n_Y}\chi^2_{\alpha}} \\ &= \sqrt{\frac{1}{n}\frac{n_{\overline{X}}}{n-n_{\overline{X}}}\frac{n-n_{\overline{Y}}}{n_{\overline{Y}}}\chi^2_{\alpha}} \\ &= \mathrm{M}_{GK}^{\alpha}(\overline{Y} \to \overline{X}). \end{aligned}$$

Par contre, si X défavorise Y, alors X favorise $\overline{Y}$ et on peut appliquer la première propriété :

$$\begin{aligned} \mathrm{M}_{GK}^{\alpha}(X \to \overline{Y}) &= \mathrm{M}_{GK}^{\alpha}(\overline{\overline{Y}} \to \overline{X}) \\ &= \mathrm{M}_{GK}^{\alpha}(Y \to \overline{X}). \end{aligned}$$

□

Corollaire 4.1. *Pour tout couple de motifs X, Y, si X favorise Y, alors les deux règles $X \to \overline{Y}$ et $\overline{Y} \to \overline{X}$ ont la même mesure selon M_{GK} et même valeur critique. Par conséquent, si les motifs $\overline{X}$ et $\overline{Y}$ sont fréquents, nous avons les équivalences, au sens de M_{GK} :*

$$\begin{aligned} X \to Y \text{ valide} &\Leftrightarrow \overline{Y} \to \overline{X} \text{ valide}, \\ X \to \overline{Y} \text{ valide} &\Leftrightarrow Y \to \overline{X} \text{ valide}. \end{aligned}$$

Selon le corollaire 4.1, on peut déduire la validité et la mesure de $\overline{Y} \to \overline{X}$ (respectivement de $Y \to \overline{X}$) à partir de celles de $X \to Y$ (respectivement de $X \to \overline{Y}$). Il est donc inutile de mettre $\overline{Y} \to \overline{X}$ ainsi que $Y \to \overline{X}$ lorsqu'on a déjà $X \to Y$ et $X \to \overline{Y}$. Contrairement aux anciennes bases des règles M_{GK}-valides, ces deux types de règles ($\overline{Y} \to \overline{X}$ et $Y \to \overline{X}$) ne feront pas partie des nouvelles bases des règles M_{GK}-valides. Cette restriction nous permet de réduire considérablement la taille des bases des règles tout en conservant les traces de toutes les informations valides.

4. Probabilité de ne pas se tromper

4.3 Choix des prémisses et des conséquents

Obtenir des connaissances fiables et exploitables à partir d'une base des données constitue la principale raison d'être de l'Extraction des Connaissances à partir des Données (ECD). Dans le cas particulier de l'extraction des règles d'association, la forme des prémisses et celle des conséquents jouent des rôles non négligeables dans l'exploitation des règles. De plus, compte tenu de la possibilité de générer un nombre trop important des règles valides, les bases des règles d'association ne doivent comprendre que des règles les plus informatives. Autrement dit, relativement à un ensemble d'axiome d'inférence, il est impératif qu'aucune redondance ne soit observée dans une base et que toutes les règles pertinentes puissent être déduites à partir des éléments constituant les bases. Compte tenu de la définition 3.8 des règles non redondantes donnée dans le chapitre 3, il est très important de bien choisir la prémisse et le conséquent des règles qui vont faire partie des bases des règles. Il faut, par exemple, s'assurer que la prémisse et le conséquent soient disjoints. En effet, la définition même d'une règle d'association précise que l'intersection de la prémisse du conséquent soit vide. Toutefois, notons qu'une bonne partie des méthodes d'extractions des bases des règles font intervenir les motifs fermés. Pourtant, entre eux, les motifs fermés peuvent avoir des intersections non vides. Dans les nouvelles bases des règles M_{GK}-valides, nous ferons en sorte que cette propriété soit respectée dans la mesure du possible. Nous allons voir justement dans le paragraphe suivant les diverses caractéristiques des prémisses et conséquents des nouvelles bases des règles M_{GK}-valides.

4.3.1 Choix de prémisse

Considérons un contexte binaire d'extraction $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ et deux motifs X et Y de $\mathscr{P}(\mathcal{I})$ [5]. Supposons qu'une règle $X \rightarrow Y$ soit valide et essayons de l'interpréter. Dans les objets (les transactions) du contexte $\mathcal{K}$, la présence des items constituant le motif X est souvent, sinon toujours (dans le cas des règles exactes) accompagnée de la présence des items constituant le motif Y. Ainsi, la présence du motif X peut être interprétée comme une condition suffisante de la présence simultanée des items constituant le motif Y. Autrement dit, dès que l'ensemble des items constituant le motif X figurent dans une transaction (un objet) du contexte $\mathcal{K}$, on y observe toujours, ou au moins, on a une très forte chance d'y observer l'ensemble des items constituant le motif Y. Par rapport à ces interprétations, si on devait répondre à la question « Quelles sont les conditions suffisantes pour qu'un motif (un conséquent) soit présent dans une transaction ? », en présence de deux ou plusieurs conditions suffisantes, il est normal et naturel de chercher seulement le strict nécessaire (les conditions les moins contraignantes, étant donnée que chacune d'elle est suffisante), pas plus. Ces arguments expliquent le fait que dans plusieurs travaux, notamment dans ceux du Pasquier, Gasmi *et al.*, Hamrouni *et al.*, les auteurs ont rapporté que pour avoir une règle plus informative, il faut que la prémisse soit minimale (au sens de l'inclusion). Rappelons que si G_X est un générateur d'un motif X, on ne peut plus trouver un motif de taille plus petit que celle de G_X dans la classe de X. Ainsi, au lieu de choisir les prémisses dans l'ensemble des motifs fermés, comme ce fut le cas dans l'ancienne base des règles approximatives M_{GK}-valides, ou dans l'ensemble des motifs pseudo-fermés comme dans la base des règles positives exactes M_{GK}-valides, nous proposons, comme dans bon nombre des bases des règles Confiance-valides, de choisir les

5. $\mathscr{P}(\mathcal{I})$ désigne l'ensemble des parties de $\mathcal{I}$.

prémisses, positives ou négatives [6], dans l'ensemble des générateurs minimaux. Ce choix va nous permettre d'avoir des règles à prémisses minimales, donc des règles plus informatives.

4.3.2 Choix du conséquent d'une règle positive

Considérons un contexte d'extraction $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$, $r_1 : X \to Y$ et $r_2 : X \to Z$ deux règles valides de ce contexte telles que $Y \subset Z$. Supposons que l'on se demande « quels sont les items présents dans la plupart ou dans la totalité des objets du contexte $\mathcal{K}$ si les items constituant le motif X sont présents dans ces objets ? » Tenant compte de la validité des règles r_1 et r_2, on obtient deux réponses possibles :

1. si les items constituant le motif X figurent dans un objet du contexte, alors on observe souvent, sinon toujours, la présence des items constituant le motif Y dans le même objet,
2. si les items constitutifs du motif X figurent dans un objet du contexte, alors on observe souvent, sinon toujours, la présence des items constitutifs du motif Z dans le même objet.

« Le motif Y inclus dans le motif Z » signifie que tous les items constituant Y sont aussi des éléments du motif Z. Il est donc clair que l'information fournie par la deuxième réponse est plus significative (dans le sens où la première réponse est incluse dans la deuxième). On dit que la règle r_1 est moins informative que la règle r_2. Une conséquence immédiate de ce constat est que, dans le cas des règles positives, les plus informatives sont celles qui ont les conséquents maximaux. Or, d'après la propriété 3.4, un motif fermé est maximal dans l'ensemble des motifs ayant la même fermeture. Donc, pour avoir une règle positive exacte ou approximative plus informative, on doit choisir le conséquent dans l'ensemble des motifs fermés.

4.3.3 Choix du conséquent d'une règle négative

Avec la mesure de qualité M_{GK}, on peut définir et extraire une base des règles négatives. Rappelons d'abord qu'une règle r est appelée règle négative lorsqu'au moins l'un de ses composants (prémisse ou conséquent) est un motif négatif. Nous étudions essentiellement les règles négatives de formes : $r_1 : X \to \overline{Y}$ et $r_2 : \overline{X} \to Y$. En effet, les règles négatives bilatérales ($\overline{X} \to \overline{Y}$) peuvent être déduites des règles positives valides via la propriété 3.6. Voyons maintenant comment choisir un conséquent d'une règle négative.
Considérons le contexte d'extraction $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ tel que $\mathcal{O} = \{o_1, o_2, ..., o_n\}$ désigne un ensemble des n objets et $\mathcal{I} = \{i_1, i_2, ..., i_m\}$ un ensemble de m-items. Soient X, Y et Z trois parties de $\mathcal{I}$. Par définition, si le motif Y modélise **la conjonction** de présence des items constituant Y dans l'ensemble des transactions, $\overline{Y}$ quant à lui, représente **la disjonction** des absences des items constituant Y.
Formellement, si on pose $Y = i_1 i_2 ... i_p$, avec $p \in \{1, 2, ..., m\}$ et $i_p \in \mathcal{I}$, on a pour chaque objet o_k de $\mathcal{O}$, $\overline{Y}(o_k) = \bigvee_{l=1}^{p} \overline{i_l}(o_k)$ où $\vee$ représente la disjonction logique. Rappelons que pour tout item i_l ($l \in \{1, ..., m\}$) et pour tout objet ou transaction o_k ($k \in \{1, ..., n\}$),

6. Une prémisse négative est un motif intervenant dans une règle négative de type : $\overline{X} \to Y$ ou $\overline{X} \to \overline{Y}$.

$$i_l(o_k) = \begin{cases} 1 \text{ si la transaction } o_k \text{ contient l'item } i_l, \\ 0 \text{ sinon.} \end{cases}$$

Par négation :

$$\bar{i}_l(o_k) = \begin{cases} 1 \text{ si la transaction } o_k \text{ ne contient pas l'item } i_l, \\ 0 \text{ sinon.} \end{cases}$$

Proposition 4.2. *Soit o_k un objet du contexte $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ et $Y, Z \in \mathscr{P}(\mathcal{I})$. Si $\overline{Y}(o_k) = 1$, alors pour tout $Z \supset Y$, $\overline{Z}(o_k) = 1$.*

Preuve.
Puisque $Z \supset Y$, il existe $T \in \mathscr{P}(\mathcal{I})$ tel que $Z = Y \cup T$. On a donc :

$$\begin{aligned} \overline{Z}(o_k) &= \overline{Y \cup T}(o_k) \\ &= \overline{Y}(o_k) \vee \overline{T}(o_k). \end{aligned}$$

Si $\overline{T}(o_k) = 1$, alors $\overline{Z}(o_k) = \overline{Y}(o_k) \vee \overline{T}(o_k) = 1 \vee 1 = 1$.
Si $\overline{T}(o_k) = 0$, alors $\overline{Z}(o_k) = \overline{Y}(o_k) \vee \overline{T}(o_k) = 1 \vee 0 = 1$ Dans tous les cas, dès que $\overline{Y}(o_k) = 1$, on a toujours : $\overline{Z}(o_k) = 1$, pour tout $Z \supset Y$. □

C'est à dire que si un motif Y est absent d'une transaction ou d'un objet o_k, ($Y(o_k) = 0$, qui est équivalent à $\overline{Y}(o_k) = 1$) alors tout motif Z contenant Y est aussi absent de la transaction (ce qui est intuitivement vrai). Donc, si les deux règles négatives à droites $r_1 : x \to \overline{Y}$ et $r_2 : x \to \overline{Z}$ sont valides, alors r_2 est moins informative que r_1.
Ce constat nous permet de comprendre que si on doit choisir entre deux règles de conséquents négatifs comparables, il vaut mieux prendre celle de conséquent minimal. D'ailleurs, selon Rioult *et al.* (2010), une règle *généralisée* $X \to \vee Y$ (avec $Y = i_1 i_2 ... i_p$ et $\vee Y = \bigvee_{l=1}^{p} i_l$) est non redondante lorsqu'elle a **un conséquent et une prémisse** minimaux. Par rapport à ces propriétés, les conséquents des règles qui constitueront les bases des règles négatives doivent être minimaux, il seront donc choisis dans l'ensemble des générateurs minimaux. Donc, en tenant compte de l'informativité des règles à prémisse minimale et la suggestion sur le choix de conséquent des règles négatives, la prémisse et le conséquent de ces dernières seront tous les deux choisis parmi des générateurs minimaux. C'est ainsi que nous avons pu proposer dans (Ramanantsoa et Totohasina, 2015) des nouvelles bases des règles M_{GK}-valides.

4.4 Nouvelle Base Positive Exacte ($NBPE$)

Avant de décrire la nouvelle base des règles positives exactes selon la mesure M_{GK}, rappelons d'abord celle qui est définie dans (Feno, 2007) :

$$BPE = \{X \to \gamma(X) \setminus X \ : \ X \text{ est } \gamma - \text{critique}\}.$$

Un motif γ- critique n'est autre qu'un motif pseudo-fermé selon l'opérateur de fermeture γ. Nous avons vu que dans BPE les prémisses ne sont pas minimales et par la suite, les conséquents ne seront pas maximaux.

Proposition 4.3. *Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte binaire d'extraction. Pour tous motifs positivement dépendants X, Y de $\mathscr{P}(\mathcal{I})$, on a l'équivalence :*

$$Conf(X \rightarrow Y \setminus X) = 1 \Leftrightarrow \mathrm{M}_{GK}(X \rightarrow Y \setminus X) = 1.$$

Preuve.
Soit X et Y deux éléments de $\mathscr{P}(\mathcal{I})$. Supposons que Conf$(X \rightarrow Y \setminus X)$=1 :

$$\begin{aligned} M_{GK}\left(X \rightarrow Y \backslash X\right) &= \frac{P\left(Y' \backslash X'\right) - P\left(Y'\right)}{1 - P\left(Y'\right)} \\ &= \frac{\mathrm{Conf}\left(X \rightarrow Y \backslash X\right) - P\left(Y'\right)}{1 - P\left(Y'\right)} \\ &= \frac{1 - P\left(Y'\right)}{1 - P\left(Y'\right)} (\text{puisque } \mathrm{Conf}(X \rightarrow Y \setminus X) = 1) \\ &= 1. \end{aligned}$$

Réciproquement, supposons que $M_{GK}\left(X \rightarrow Y \backslash X\right) = 1$,

$$\begin{aligned} \text{ce qui nous donne } \frac{P\left(Y' \backslash X'\right) - P\left(Y'\right)}{1 - P\left(Y'\right)} &= 1, \\ \text{soit } \quad P\left(Y' \backslash X'\right) - P\left(Y'\right) &= 1 - P\left(Y'\right), \end{aligned}$$

donc $P\left(Y' \backslash X'\right) = 1$, d'où l'égalité Conf$(X \rightarrow Y \setminus X) = 1$. □

Constatant l'équivalence entre règle exacte confiance-valide et règle exacte M_{GK}-valide, au lieu de prendre la base de Duquenne-Guiges comme base des règles M_{GK}-valides, on peut plutôt prendre la base générique des règles exactes définie dans (Pasquier, 2000a). Autrement dit, l'équivalence (4.3) et le souhait d'avoir une prémisse minimale nous permet de prendre la base générique des règles exactes (BG) comme base des règles positives exactes M_{GK}-valides. En effet, dans BG, les prémisses sont choisies dans l'ensemble des générateurs, donc elles sont minimales et les conséquents sont choisis dans l'ensemble des motifs fermés, donc ils sont forcément maximaux. Donc, la définition de nouvelle base des règles exactes M_{GK}-valides ($NBPE$) doit tenir compte de ces points.
Précisons que les règles exactes constituant la base générique BG sont obtenues à partir de deux motifs (prémisse et conséquent) d'une même classe, c'est-à-dire deux motifs de même fermeture. Nous allons voir, à travers la proposition 4.4 qu'avec la mesure M_{GK}, les règles positives exactes sont obtenues de la même manière.

Proposition 4.4. *Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte d'extraction. Pour tous motifs X, Y de $\mathscr{P}(\mathcal{I})$ tels que $\gamma(X) = \gamma(Y) = F$, on a toujours :* $\begin{cases} Supp(Y \setminus X) \geq Supp(Y), \\ M_{GK}\left(X \rightarrow Y \setminus X\right) = 1, \\ Supp\left(X\right) = Supp\left(F\right). \end{cases}$

Preuve.
En se servant de la propriété du support d'un motif, comme $Y \setminus X \subseteq Y$, on a toujours :

$$\mathrm{Supp}(Y \setminus X) \geq \mathrm{Supp}(Y).$$

D'après la propriété de fermeture, on a : pour tout motif X, $\text{Supp}(X) = \text{Supp}(\gamma(X))$ et donc égal au support de F.
Soit X et Y deux motifs tels que : $\gamma(X) = \gamma(Y) = F$:

$$\begin{aligned}
M_{GK}(X \to Y \setminus X) &= \frac{\frac{\text{Supp}(X \cup Y)}{\text{Supp}(X)} - \text{Supp}(Y \setminus X)}{1 - \text{Supp}(Y \setminus X)} \\
&= \frac{\frac{\text{Supp}(\gamma(X \cup Y))}{\text{Supp}(\gamma(X))} - \text{Supp}(Y \setminus X)}{1 - \text{Supp}(Y \setminus X)} \\
&= \frac{\frac{\text{Supp}(\gamma(\gamma(X) \cup \gamma(Y)))}{\text{Supp}(\gamma(X))} - \text{Supp}(Y \setminus X)}{1 - \text{Supp}(Y \setminus X)} \\
&= \frac{\frac{\text{Supp}(F)}{\text{Supp}(F)} - \text{Supp}(Y \setminus X)}{1 - \text{Supp}(Y \setminus X)} \; (\gamma(X) = \gamma(Y) = \gamma(F) = F) \\
&= \frac{1 - \text{Supp}(Y \setminus X)}{1 - \text{Supp}(Y \setminus X)} \\
&= 1.
\end{aligned}$$

□

Une conséquence immédiate de la proposition 4.4 nous donne l'axiome d'inférence (RPE).

Corollaire 4.2 (Axiome d'inférence (RPE))**.**
Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte binaire d'extraction. Soient F un fermé de $\mathscr{P}(\mathcal{I})$ et G un générateur de F. Si $G \to F \setminus G$ est une règle exacte valide, alors pour tous motifs Z, T de $[F]$, la règle $Z \to T \setminus Z$ est une règle exacte valide.

À partir de ce corollaire, nous allons proposer une nouvelle base des règles positives exactes. On rappelle que l'extraction de base des règles a pour objectif de sélectionner un ensemble minimal des règles à partir duquel on peut dériver les autres règles via un ensemble d'axiome d'inférence approprié. Cette réduction doit se faire sans perte d'information, c'est-à-dire que toutes les règles valides peuvent être retrouvées avec leurs mesures (Support et M_{GK} dans notre cas) et toutes les règles dérivées sont valides.

Proposition 4.5. *Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte binaire d'extraction. Désignons par $FF_{\mathcal{K}}$ l'ensemble des fermés fréquents et par $\mathcal{G}_F$ l'ensemble des générateurs d'un motif fermé F. L'ensemble $NBPE$ défini par :*

$$NBPE = \{r : G \to F \setminus G \text{ telle que } F \in FF_{\mathcal{K}}, G \in \mathcal{G}_F \text{ et } G \neq F\}. \tag{4.5}$$

est une base pour les règles positives exactes M_{GK}-valides relativement à l'axiome d'inférence (RPE).

Preuve.
Montrons que l'ensemble $NBPE$ est un ensemble minimal à partir duquel on peut déduire l'ensemble des règles exactes valides via l'axiome d'inférence RPE.
Prenons deux motifs distincts X et Y tels que $X \subset Y$ et la règle $r : X \to Y \setminus X$ soit une règle exacte valide. Montrons que : soit r est dans $NBPE$, soit elle peut être déduite d'une

règle dans $NBPE$ par l'application de l'axiome (RPE).

$X \to Y \setminus X$ exacte valide signifie : $\begin{cases} M_{GK}(X \to Y \setminus X) = 1, \\ \text{Supp}(X) \geq minSupp, \\ \text{Supp}(Y \setminus X) \geq minSupp. \end{cases}$.

Selon la proposition 4.3, $M_{GK}(X \to Y \setminus X) = 1 \Leftrightarrow \text{Conf}(X \to Y \backslash X) = 1$.

Or Conf $(X \to Y \backslash X) = 1 \Leftrightarrow \dfrac{\text{Supp}(X \cup Y)}{\text{Supp}(X)} = 1$, donc $\text{Supp}(X \cup Y) = \text{Supp}(X)$. Comme $X \subset Y$, on a toujours l'égalité $X \cup Y = Y$. Autrement dit, Conf$(X \to Y \setminus X)$=1 est équivalent à Supp(X)=Supp(Y). D'autre part, si $X \subset Y$ alors $\phi(X) \supseteq \phi(Y)$, où ϕ désigne l'application qui associe un motif à son extension. De plus, $\text{Supp}(X) = \text{Supp}(Y)$ est équivalent à $|\phi(X)| = |\phi(Y)|$. À partir de $\phi(X) \supseteq \phi(Y)$ et $|\phi(X)| = |\phi(Y)|$ on peut déduire que $\phi(X) = \phi(Y)$, par conséquent $\psi \circ \phi(X) = \psi \circ \phi(Y)$ où ψ désigne l'application qui associe un objet à son intension. D'où l'implication :

$$X \subset Y \text{ et } M_{GK}(X \to Y \backslash X) = 1 \Rightarrow \gamma(X) = \gamma(Y) = F.$$

Plus précisément, si $X \to Y \setminus X$ est exacte et valide, alors il existe un fermé fréquent F tel que $\gamma(X) = \gamma(Y) = F$. Deux cas sont possibles : X générateur et Y fermé et le cas contraire. Dans le premier cas, la règle $X \to Y \setminus X$ est dans $NBPE$; dans le cas contraire, il existe un fermé F et un générateur G_F de F tels que $\gamma(G_F) = \gamma(X) = \gamma(Y) = \gamma(F)$ et la règle $G_F \to F \setminus G_F$ sera dans $NBPE$. Dans ce cas, on peut appliquer (RPE) à la règle $G_F \to F \setminus G_F$ pour retrouver la règle $X \to Y \setminus X$ puisque pour tout motif X de l'ensemble $[F]$, on peut toujours trouver un générateur G_F de F tel que $G_F \subseteq X$ (l'inclusion est stricte lorsque X n'est pas un générateur) et pour tout motif Y de $[F]$, $Y \subseteq F$ car, F est un fermé, donc maximal au sens de l'inclusion.

Supposons maintenant que X et Y ne sont pas comparables (on n'a ni $X \subset Y$, ni $Y \subset X$). Remarquons que pour tous X et Y de $\mathscr{P}(\mathcal{I})$, $X \cup Y$ est dans $\mathscr{P}(\mathcal{I})$ et de plus $Y \setminus X$ et $(X \cup Y) \setminus X$ désignent un même et unique motif. Donc, $X \to Y \setminus X$ et $X \to Z \setminus X$ avec $Z = X \cup Y$ désignent une même et unique règle. Si $X \to Y \setminus X$ est M$_{GK}$-valide, alors il existe toujours une règle $X \to Z \setminus X$ avec $X \subset Z$ et $X \to Z \setminus X$ coïncide avec $X \to Y \setminus X$. Le précédent raisonnement peut être appliqué à la règle $X \to Z \setminus X$ pour montrer que soit cette règle est élément de $NBPE$ soit elle peut être déduite d'une règle de $NBPE$ par l'application de l'axiome d'inférence RPE.

Montrons maintenant qu'une règle dérivée par (RPE) est toujours une règle valide. Prenons une règle $r : X \to Y \setminus X$ dans $NBPE$ et désignons par $r_1 : X_1 \to Y_1 \setminus X_1$ une règle dérivée de r par l'application de (RPE).

$$\text{La règle } r : X \to Y \setminus X \text{ dans } NBPE \text{ signifie que } \begin{cases} X \text{ générateur et } Y \text{ fermé}, \\ M_{GK}(X \to Y \backslash X) = 1, \\ \text{Supp}(X) \geq \text{minSupp}, \\ \text{Supp}(Y \backslash X) \geq \text{minSupp}. \end{cases}$$

La règle $r_1 : X_1 \to Y_1 \setminus X_1$ est une règle dérivée de r par l'application de l'axiome (RPE) nous permet d'affirmer que :

$$X_1 \supseteq X,\ Y_1 \subseteq Y \text{ et } \begin{cases} \gamma(X_1) = \gamma(X) \\ \gamma(Y_1) = \gamma(Y). \end{cases}$$

Or $\gamma(X) = \gamma(Y) = Y$, donc X_1 et Y_1 sont dans la classe de Y, c'est-à-dire que X_1 et Y_1 dans une même classe et X_1 et $Y_1 \setminus X_1$ sont positivement dépendants. En se servant de la proposition 4.4, on a :

$$\begin{cases} M_{GK}\left(X_1 \rightarrow Y_1 \backslash X_1\right) = 1, \\ \text{Supp}(X_1) = \text{Supp}(\gamma(X_1)) = \text{Supp}(X), \\ \text{Supp}\left(Y_1 \backslash X_1\right) = \text{Supp}\left(\gamma\left(Y_1\right)\right) = \text{Supp}\left(Y\right). \end{cases}$$

La règle $X_1 \rightarrow Y_1 \setminus X_1$ est une règle exacte et valide. Donc, l'axiome d'inférence (RPE) est valide et complet. De plus, on peut déterminer exactement les mesures des règles dérivées.
Montrons maintenant que l'ensemble $NBPE$ est minimal. Soient $\mathcal{K}$ un contexte binaire d'extraction et $NBPE$ la nouvelle base des règles positives exactes associées.
Soit $r : X \rightarrow Y \setminus X$ un élément de $NBPE$. Désignons par $NBPE'$ l'ensemble $NBPE \setminus \{r : X \rightarrow Y \setminus X\}$. Montrons qu'il est impossible d'engendrer r à partir des éléments de $NBPE'$ par l'utilisation de l'axiome d'inférence (RPE).
D'abord, $r : X \rightarrow Y \backslash X \in NBPE$ signifie que X est un générateur et Y un fermé. Effectuons un raisonnement par l'absurde. Supposons qu'il existe une règle $r_1 : X_1 \rightarrow Y_1 \backslash X_1$ de $NBPE'$ telle que l'application de (RPE) à r_1 engendre la règle $r : X \rightarrow Y \setminus X$. Cela signifie que :

$$\begin{cases} X \supset X_1 \text{ et } \gamma\left(X\right) = \gamma\left(X_1\right) \\ Y \subset Y_1 \text{ et } \gamma\left(Y\right) = \gamma\left(Y_1\right) \end{cases}$$

Comme r_1 est un élément de $NBPE'$, $\gamma(X_1) = \gamma(Y_1)$. Autrement dit, X, X_1, Y et Y_1 sont dans une même classe (dans $[Y]$). Donc X ne peut pas être un générateur (puisqu'il existe X_1 dans sa classe tel que $X \supset X_1$) et Y ne peut pas être un fermé (puisque qu'il existe Y_1 dans sa classe tel que $Y \subset Y_1$). Donc, la règle $r : X \rightarrow Y \setminus X$ ne peut pas être un élément de $NBPE$. ceci est en contradiction avec la supposition de départ. Cela montre que l'ensemble $NBPE$ est un ensemble minimal des règles à partir duquel on peut engendrer les autres règles exactes valides via l'axiome d'inférence (RPE). □

4.4.1 Exemple

Remarquons d'abord que si l'ancienne base des règles exactes selon la mesure M_{GK} définie dans (Feno, 2007) coïncide avec la base de Duquenne-Guigues, la nouvelle base des règles positives exactes selon la mesure M_{GK} $(NBPE)$ quant à elle, elle coïncide avec la base générique des règles exactes définie dans (Pasquier, 2000a). Pour un $minSupp = 2/9$, si on considère le contexte $\mathcal{K}$ du tableau 2.1, nous obtenons les bases du tableau 4.4.

Base de Duquenne-Guigues	Supports
$E \rightarrow B$	2/3
$C \rightarrow AB$	1/2
$F \rightarrow ABE$	1/3

$NBPE$	Support
$E \rightarrow B$	2/3
$AE \rightarrow B$	1/2
$DE \rightarrow B$	1/3
$C \rightarrow AB$	1/2
$F \rightarrow ABE$	1/3
$CE \rightarrow AB$	1/3

Tableau 4.4 – Base de Duquenne-Guigues et $NBPE$ avec un $minSupp = 1/3$

Remarque On peut tout de suite remarquer que la nouvelle base contient beaucoup plus de règles que l'ancienne et que l'on pourrait croire qu'aucune amélioration n'a été obtenue. Comme l'a souligné Pasquier (2000b), la base générique qui contient beaucoup plus de règles que la base de Duquenne-Guigues est la base minimale sans perte d'information pour les règles d'association exactes. En effet, si toutes les règles exactes sont déductibles à partir de la base de Duquenne-Guigues, il n'en est pas de leurs supports. De plus, contrairement à la base générique, donc à la nouvelle base des règles positives exactes M_{GK}-valides, la base de Duquenne-Guigues n'est pas constituée des règles à prémisse minimale et conséquent maximal. À titre d'exemple, si on avait pris un $minSup = 1/6$, on aurait pu avoir le motif $ABDE$ comme pseudo-fermé fréquent. Par conséquent, on aurait pu mettre dans la base la règle $r_1 : ABDE \rightarrow F$. Pourtant, DF est le générateur du motif fermé $ABDEF$ (fermeture de $ABDE$). Donc, à la place de la règle $r_1 : ABDE \rightarrow F$ de la base de Duquenne-Guigues (donc, dans l'ancienne base des règles exactes M_{GK} valides), on a, dans la base générique (donc, dans la nouvelle base des règles exactes M_{GK} valides) la règle $r_2 : DF \rightarrow ABE$. On voit que r_2 est de loin plus informative que r_1. Nous allons voir qu'avec cette nouvelle base, non seulement on peut retrouver toutes les règles exactes, mais on peut aussi retrouver leurs supports.

4.5 Nouvelle Base Négative Exacte ($NBNE$)

Compte tenu des critiques sur l'informativité des règles constituant l'ancienne base des règles négatives exacte, dans cette section, nous allons proposer une nouvelle description de la base des règles négatives exactes.
Pour avoir une règle exacte valide de type $X \rightarrow \overline{Y}$, il faut que l'on ait :

$$\begin{cases} M_{GK}\left(X \rightarrow \overline{Y}\right) = 1, \\ Supp\left(\overline{Y}\right) \geq minSupp, \\ Supp\left(X\right) \geq minSupp. \end{cases}$$

Nous avons montré qu'il y a une équivalence entre $\mathrm{M}_{GK}(X \rightarrow \overline{Y}) = 1$ et $\mathrm{Supp}(X \rightarrow Y) = 0$. C'est pour cette raison que dans l'ancienne BNE, le fait de choisir X dans $Bd^{+}(0)$ est très judicieux. Avant de définir une nouvelle base négative exacte, nous allons rappeler sous forme de lemme la description de l'ancienne base des règles négatives exactes.

Lemme 4.1 ((Feno *et al.*, 2006 ; Feno, 2007))**.**
L'ensemble $BNE = \{X \rightarrow \{\overline{x}\} : X \in Bd^{+}(0)$ *et* $x \notin X\}$ *est une base des règles négatives exactes relativement aux axiomes d'inférences donnés ci-après :*

$$(RNE1) \quad Si \begin{cases} M_{GK}\left(X \rightarrow \overline{Y}\right) = 1 \\ Supp\left(Y \cup Z\right) > 0, \end{cases} \quad alors\ M_{GK}\left(X \rightarrow \overline{Y \cup Z}\right) = 1.$$

$$(RNE2) \quad Si \begin{cases} M_{GK}\left(X \rightarrow \overline{Y}\right) = 1 \\ Z \subset X \\ Supp\left(Z \cup Y\right) = 0, \end{cases} \quad alors\ M_{GK}\left(Z \rightarrow \overline{Y}\right) = 1.$$

Remarquons que les éléments de $Bd^{+}(0)$ sont des fermés maximaux. On sait pourtant que pour un générateur G_X de X, $M_{GK}\left(X \rightarrow \overline{Y}\right)$ est égal à $M_{GK}\left(G_X \rightarrow \overline{Y}\right)$. Donc, au lieu de

mettre la règle $r_1 : X \rightarrow \overline{Y}$ dans la base BNE, nous pouvons plutôt considérer la règle $r_2 : G_X \rightarrow \overline{Y}$, elle est plus informative que r_1 (puisque $G_X \subset X$).

Proposition 4.6. *L'ensemble $NBNE$ défini par :*

$$NBNE = \left\{ \begin{array}{l} G_X \rightarrow \{\overline{x}\} : \ X \in Bd^+(0), G_X \in \mathcal{G}_X, \ x \notin X \\ et \ \min(Supp(X), Supp(\{\overline{x}\})) \geq minSupp \end{array} \right\}$$

est une base des règles négatives exactes relativement aux axiomes d'inférence $(RNE1)$ et $(RNE2)$. $\mathcal{G}_X$ désigne l'ensemble des générateurs de X.

Preuve. Définissons une relation R qui associe une règle $r : G_X \rightarrow \{\overline{x}\}$ de la $NBNE$ à la règle $r' : \gamma(G_X) \rightarrow \{\overline{x}\}$ de BNE. La relation R est une application surjective et non injective de $NBNE$ dans BNE.
En effet, pour chaque générateur G_X, le motif $\gamma(G_X)$ est unique et il est fermé. Donc, pour chaque règle $r : G_X \rightarrow \{\overline{x}\}$ de $NBNE$, il existe une unique règle $r' : \gamma(G_X) \rightarrow \{\overline{x}\}$ de BNE telle que $r' = R(r)$. Nous savons aussi que chaque motif fermé a au moins un générateur. Donc, pour chaque élément $r' : X \rightarrow \{\overline{x}\}$ de BNE (X est un élément de $Bd^+(0)$ du contexte étudié), il existe au moins une règle r de $NBNE$ telle que $r' = R(r)$. La relation R est donc une application surjective. Par conséquent, pour chaque règle r de $NBNE$, il existe une unique règle r' de BNE telle que $r' = R(r)$, donc, $R(NBNE) = BNE$. L'utilisation de l'application R et le lemme 4.1 montre que la nouvelle bas $NBNE$ est génératrice.
Étudions maintenant la minimalité de $NBNE$.
Un motif fermé X peut avoir plusieurs générateurs. Supposons que G_{1X} et G_{2X} sont des générateurs d'un fermé X. Dans ce cas :

$$R(G_{1X} \rightarrow \{\overline{x}\}) = R(G_{2X} \rightarrow \{\overline{x}\}) = X \rightarrow \{\overline{x}\}.$$

Donc, l'application R n'est pas forcement injective. Cela signifie que $Card(NBNE)$ est plus grand ou égal à $Card(BNE)$. Nous aurons pu construire $NBNE$ en prenant un seul générateur de chacun des motifs fermés et ceci aurait été suffisant pour trouver l'ensemble des règles négatives exactes valides puisque dans le cas échéant, R est bijective et l'utilisation de $RNE1$ et $RNE2$ avec les éléments de $R(NBNE)$ (qui n'est autre que BNE) aurait donnée toutes les règles négatives exactes valides. Dans ce cas, selon le lemme 4.1, $NBNE$ aurait été minimale et génératrice. Mais, puisqu'il est très rare de dériver et d'interpréter toutes les règles exactes valides et de plus, les générateurs d'un fermé quelconque ne sons jamais comparables entre eux, nous avons jugé utile de mettre dans la base les règles issues de tous les générateurs d'un élément de $Bd^+(0)$ quelconque. Dans ce cas, la taille de $NBNE$ est supérieure ou égale à celle de BNE, et par conséquent, $NBNE$ ne sera plus minimale mais la nouvelle base des règles négatives est plus informative (les prémisses sont minimaux et il n'est pas nécessaire d'utiliser les axiomes d'inférence pour avoir toutes les information capitale) que l'ancienne base des règles négatives exactes. En conclusion, par le fait qu'un générateur d'un fermé n'est pas forcement unique, $NBNE$ n'est pas toujours minimale, mais elle est génératrice et plus informative que $NBNE$. □

Exemple 9. *Pour le contexte décrit dans le tableau 2.1, nous avons :*

$$Bd^+(0) = \{ABCD, \ ABCDEF, \ ABCEF\}.$$

En prenant un $minSupp = 1/6$, l'ancienne et la nouvelle base des règles négatives exactes sont données les tableaux ci-dessous (cf. Tableau 4.5) :

$Bd^+(0)$	Ancienne BNE
$ABCD$	$ABCD \rightarrow \overline{E}$ $ABCD \rightarrow \overline{F}$
$ABCEF$	$ABCEF \rightarrow \overline{D}$
$ABDEF$	$ABDEF \rightarrow \overline{C}$

$Bd^+(0)$	Nouvelle BNE
$ABCD$	$CD \rightarrow \overline{E}$ $CD \rightarrow \overline{F}$
$ABCEF$	$CF \rightarrow \overline{D}$
$ABDEF$	$DF \rightarrow \overline{C}$ $ADE \rightarrow \overline{C}$

Tableau 4.5 – Comparaison des bases des règles négatives exactes

4.6 Nouvelle Base Positive Approximative ($NBPA$)

Dans la pratique, une règle souffre souvent des exceptions et le fait d'avoir « *quelques* » exceptions ne constitue pas toujours une raison suffisante pour la rejeter. Le problème est de savoir quand est-ce que les exceptions ne seront plus acceptables et qu'on sera obligé de rejeter la règle. D'où l'importance des études sur les règles approximatives. Une base des règles approximatives valides selon la mesure M_{GK} est proposée dans (Feno, 2007). Dans le but de l'améliorer, nous allons apporter des modifications dans la description de cette base des règles. Pour mieux comprendre les modifications apportées, nous allons rappeler sous forme de lemme la description de l'ancienne base des règles positives approximatives et les trois points que l'on peut améliorer.

Lemme 4.2 ((Feno *et al.*, 2006 ; Feno, 2007)). *Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte d'extraction, α un niveau de confiance fixé par l'utilisateur et X, Y deux parties de $\mathcal{I}$. Relativement à l'axiome d'inférence (PA), l'ensemble $BPA(\alpha)$ défini par :*
$BPA(\alpha) = \{X \rightarrow Y : \gamma(X) = X, \gamma(Y) = Y \text{ et } Supp(Y)(1-\alpha) + \alpha \leq Conf(X \rightarrow Y) < 1\}$
constitue une base pour les règles positives approximatives valides selon la mesure M_{GK} au niveau de confiance α. $Conf$ désigne la mesure confiance (cf. tableau 2.2) et (PA) l'axiome d'inférence ci-après : Si $X \rightarrow Y$ est valide et Z, T deux motifs tels que $\gamma(Z) = \gamma(X)$ et $\gamma(T) = \gamma(Y)$, alors $Z \rightarrow T$ est une règle valide.

Avant de proposer une nouvelle base des règles approximatives M_{GK}-valides, rappelons brièvement les trois points à améliorer dans l'ancienne description de la base positive approximative.

- Pour valider une règle $X \rightarrow Y$, au lieu d'utiliser le niveau de confiance α, on doit plutôt utiliser la valeur critique de M_{GK} calculée au niveau de confiance α (cf. définition 4.2) pour prendre une décision sur la validité d'une règle $X \rightarrow Y$. Le paragraphe 4.2 montre la nécessité de la considération de cette valeur critique.
- Comme le motif X est fermé, si l'ensemble des motifs ayant la même fermeture X, c'est à dire, la classe de X ne se réduit pas à un singleton, alors X ne peut pas être minimal (au sens de l'inclusion) dans cet ensemble et, par la suite, le conséquent $Y \setminus X$ ne sera pas maximal non plus. Pour avoir une règle plus informative (de prémisse minimale et conséquent maximal), pour la prémisse, on peut plutôt prendre un générateur à la place d'un fermé.

- Les motifs X et Y sont des fermés, donc ils ne sont pas forcément disjoints. Pour éviter cette intersection non vide, on peut prendre des règles de type $X \rightarrow Y \setminus X$ comme éléments qui vont constituer la base des règles positives approximatives.

Ces trois points nous guident pour la définition de la nouvelle base des règles d'association positives approximatives. Avant de voir la description de la nouvelle base positive approximative valide selon la mesure M_{GK}, nous allons voir l'impact du choix des règles de type $X \rightarrow Y \setminus X$ sur les axiomes d'inférence permettant de retrouver les autres règles valides.

4.6.1 Variation de M_{GK} par rapport au support de conséquent

Nous avons souligné maintes fois que l'intersection non vide entre prémisse et conséquent est l'un des points que l'on peut améliorer de l'ancienne base des règles M_{GK}-valides, ce changement va créer un impact sur les axiomes d'inférence permettant de retrouver l'ensemble des règles valides. Par exemple, dans l'ancienne base où on a pris des règles de type $X \rightarrow Y$, avec X et Y fermés, du moment que la règle $X \rightarrow Y$ est valide, alors pour tout Z dans la classe de X (classe de X, notée souvent par $[X]$, désigne l'ensemble des motifs ayant comme fermeture le motif X) et pour tout T dans la classe de Y (voir Fig. 4.1), la règle $Z \rightarrow T$ est toujours valide. En effet, pour tous motifs Z dans $[X]$ et T dans $[Y]$, on a :

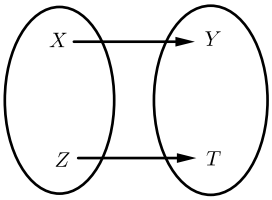


FIGURE 4.1 – Dérivation dans l'ancienne base positive approximative

$$\begin{aligned}
M_{GK}\left(Z \rightarrow T\right) &= \frac{\frac{\mathrm{Supp}(Z \cup T)}{\mathrm{Supp}(Z)} - \mathrm{Supp}\left(T\right)}{1 - \mathrm{Supp}\left(T\right)} \\
&= \frac{\frac{\mathrm{Supp}(\gamma(Z \cup T))}{\mathrm{Supp}(\gamma(Z))} - \mathrm{Supp}\left(\gamma\left(T\right)\right)}{1 - \mathrm{Supp}\left(\gamma\left(T\right)\right)} \\
&= \frac{\frac{\mathrm{Supp}(\gamma(\gamma(Z) \cup \gamma(T)))}{\mathrm{Supp}(\gamma(Z))} - \mathrm{Supp}\left(\gamma\left(T\right)\right)}{1 - \mathrm{Supp}\left(\gamma\left(T\right)\right)} \\
&= \frac{\frac{\mathrm{Supp}(\gamma(X \cup Y))}{\mathrm{Supp}(X)} - \mathrm{Supp}\left(Y\right)}{1 - \mathrm{Supp}\left(Y\right)} \\
&= \frac{\frac{\mathrm{Supp}(X \cup Y)}{\mathrm{Supp}(X)} - \mathrm{Supp}\left(Y\right)}{1 - \mathrm{Supp}\left(Y\right)} \\
&= M_{GK}\left(X \rightarrow Y\right).
\end{aligned}$$

On a donc une équivalence entre la validité de la règle $X \rightarrow Y$ et celle de $Z \rightarrow T$. Maintenant, au lieu de mettre $X \rightarrow Y$ dans la base des règles positives, pour éviter l'intersection non vide

entre les fermés, nous allons choisir les règles de type $X \to Y \setminus X$ et, examinons l'impact de ce choix sur l'axiome d'inférence (RPA). Pour cela, étudions la validité de la règle $Z \to T \setminus Z$ comparativement à celle de $X \to Y \setminus X$. D'abord, selon la propriété 3.8, pour tous motifs X, Y, il vient :

$$\text{Conf}\,(X \to Y \backslash X) = \text{Conf}\,(\gamma(X) \to \gamma(Y) \backslash \gamma(X))\,.$$

Donc, si on ne se réfère qu'à la probabilité conditionnelle (mesure Confiance) de ces deux règles, on peut très vite se rendre compte de l'équivalence entre la validité de $X \to Y \setminus X$ et celle de $Z \to T \setminus Z$, pour tous Z dans $[X]$ et T dans $[Y]$. On aurait donc pu conclure que la validité de $X \to Y \setminus X$ et celle de $Z \to T \setminus Z$ sont équivalentes et cela, pour tout Z dans $[X]$ et pour tout T dans $[Y]$. Pourtant, pour tout T dans $[Y]$, $T \subseteq Y$, donc $T \setminus Z \subseteq Y$ et $\text{Supp}(T \setminus Z) \geqslant \text{Supp}(Y)$. Donc, $X \to Y \setminus X$ et $Z \to T \setminus Z$, pour un Z dans $[X]$ et T dans $[Y]$ sont des règles de même Confiance, mais des supports de conséquent qui peuvent être différents. Il est donc nécessaire d'étudier la variation de la mesure M_{GK} par rapport à la variation de la taille des motifs conséquents. Pour tous motifs X, Y, supposons que la $\text{Conf}(X \to Y)$ est fixée à une valeur constante et étudions la variation de M_{GK} par rapport aux supports de conséquent :

$$\begin{aligned} M_{GK}\,(X \to Y) &= \frac{\text{Conf}\,(X \to Y) - \text{Supp}\,(Y)}{1 - \text{Supp}\,(Y)}, \\ \frac{\partial M_{GK}\,(X \to Y)}{\partial \text{Supp}\,(Y)} &= \frac{-1 + \text{Conf}\,(X \to Y)}{(1 - \text{Supp}\,(Y))^2} < 0. \end{aligned}$$

On voit ici que la mesure M_{GK} est une fonction décroissante des supports de conséquents. La figure 4.2 représente la composante positive de la mesure M_{GK} et le comportement de cette mesure par rapport aux règles de même Confiance et de support des conséquents comparables.

Ces courbes nous montrent que même si on se fixe une valeur de probabilité conditionnelle, comme c'est le cas de la famille des règles $X_1 \to Y_1 \setminus X_1$ avec $X_1 \in [X]$ et $Y_1 \in [Y]$, les valeurs de M_{GK} peuvent varier d'une règle à l'autre. Pour se fixer les idées, supposons qu'on ait deux motifs X et Y tels que $\text{Conf}(X \to Y \setminus X) = 0,8$ (courbe bleue). Ensuite, prenons un motif Y_1 dans $[Y]$.

$$\begin{cases} Y_1 \in [Y] \\ Y \text{ fermé} \end{cases} \Rightarrow Y_1 \subseteq Y$$

Donc, pour tout motif X, $Y_1 \backslash X \subseteq Y \backslash X$ et, par conséquent, $\text{Supp}(Y_1 \backslash X) \geqslant \text{Supp}(Y \backslash X)$ (propriété d'antimonotonie de support). Comme M_{GK} est une fonction décroissante des supports des conséquents (pour une confiance fixée), on a : $M_{GK}\,(X \to Y_1 \backslash X) \leq M_{GK}\,(X \to Y \backslash X)$. Ce constat nous permet de rendre compte que contrairement à la mesure Confiance, la validité d'une règle $X \to Y \setminus X$ ne suffit pas pour se prononcer sur la validité d'une règle $X \to Z \setminus X$ avec $Z \subset Y$ et $\gamma(Z) = \gamma(Y)$. D'ailleurs, si on représente la composante négative de M_{GK}, on se rend compte qu'au fur et à mesure de la diminution de la taille des motifs conséquents, donc de l'augmentation de support des conséquents, la valeur de M_{GK} décroit et passe par la valeur positive, s'annule et bascule dans la partie négative. Tout dépend de la place du motif $Z \setminus X$ dans les classes formées par les motifs ayant la même fermeture. Nous allons voir dans le prochain paragraphe les différents emplacements possibles de $Z \setminus X$ par rapport à $[Y]$.

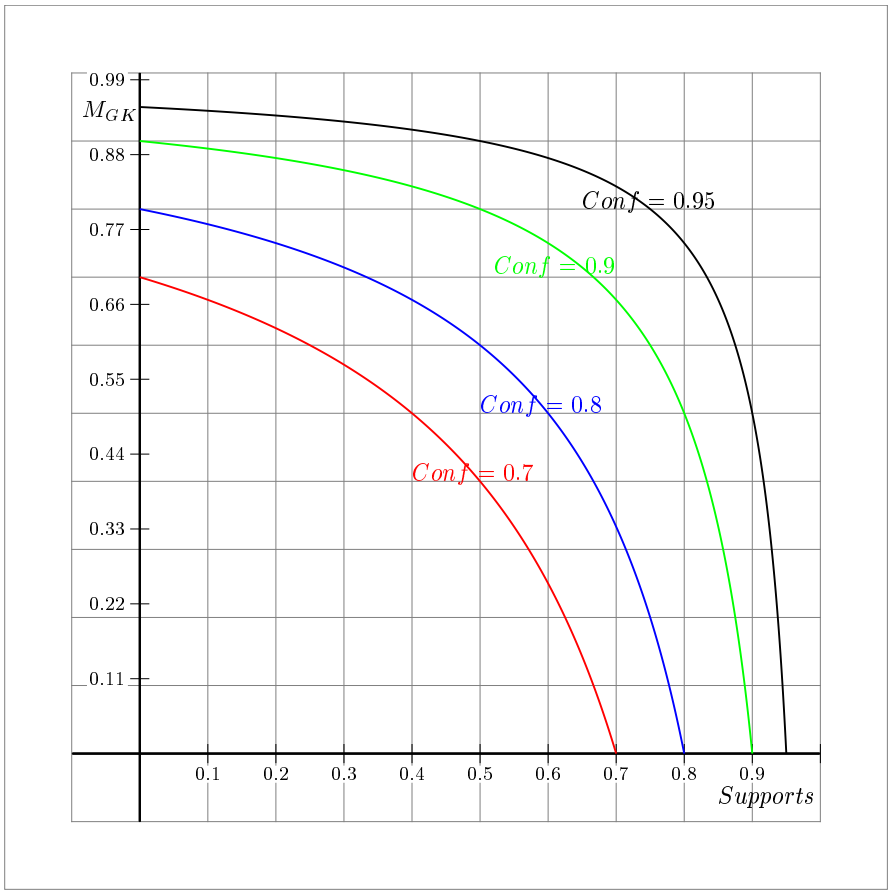


FIGURE 4.2 – M_{GK} en fonction de support de conséquent

4.6.2 Emplacement du motif $Y \setminus X_1$ par rapport à $[Y]$

Considérons un motif fermé Y et un générateur X_1 de la classe d'un autre fermé X. Si on veut étudier et éventuellement envisager de mettre la règle $X_1 \rightarrow Y \setminus X_1$ dans la nouvelle base des règles approximatives valides selon la mesure M_{GK}, il va falloir découvrir dans quelle classe des fermés se trouve le conséquent $Y \setminus X_1$. Trois cas, représentés par la figure 4.3 sont envisageables.

Le premier cas représente le cas idéal parce que le motif $Y \backslash X_1$ reste dans la classe de $[Y]$. Dans ce cas, $M_{GK}(X \rightarrow Y) = M_{GK}(X_1 \rightarrow Y \backslash X_1)$. Donc, au lieu de mettre $X \rightarrow Y$ dans la base des règles approximatives, on peut plutôt utiliser la règle $X_1 \rightarrow Y \backslash X_1$ comme représentante des règles ayant la prémisse dans $[X]$ et le conséquent dans $[Y]$. En effet, $X_1 \rightarrow Y \backslash X_1$ est formée par une prémisse minimale, un conséquent maximal dont l'intersection avec la prémisse est vide. Du point de vue sémantique, $X_1 \rightarrow Y \backslash X_1$ est plus informative que la règle $X \rightarrow Y$, donc, entre ces deux règles, il est préférable de mettre $X_1 \rightarrow Y \backslash X_1$ dans la base. Toutefois, il y a les deux autres situations à analyser. Dans le cas où le motif $Y \backslash X_1$ ne reste plus dans la classe de Y, il peut être dans la classe d'un autre fermé Z (2e Cas) ou encore, il peut être un fermé dans une autre classe (3e Cas), mais, en tout cas, on a : $M_{GK}(X \rightarrow Y) \neq M_{GK}(X_1 \rightarrow Y \backslash X_1)$. Pourtant, il n'est pas exclu que la règle $X \rightarrow Y$ et les règles dérivées de type $Z \rightarrow T \setminus Z$, avec Z dans la classe de X et $T \setminus Z$ dans la classe de

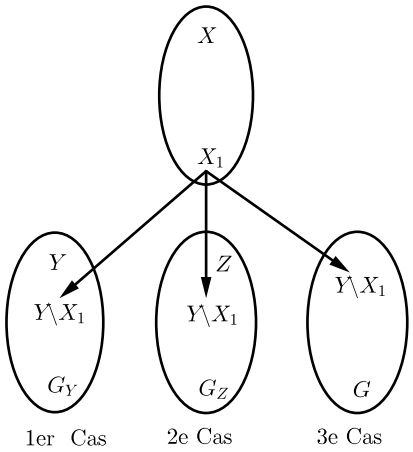


FIGURE 4.3 – Différentes possibilités de l'emplacement du motif $Y \setminus X$

Y, soient valides. Dans ce cas, puisque la validité de $X_1 \rightarrow Y \backslash X_1$ ne permet pas de dériver celle de $X \rightarrow Y$, pour ne pas perdre une famille des règles ayant une prémisse dans $[X]$ et un conséquent dans $[Y]$, il est impératif que l'on trouve une règle qui va représenter cette famille dans la base des règles. Certes, il faut qu'elle soit la plus représentative possible.
On peut, par exemple, changer du générateur ou prendre un autre motif non minimal (du côté de prémisse) ou encore, passer aux motifs non fermés (du côté de conséquent). L'objectif est de trouver un motif $X_1 \in [X]$ et un motif $Y_1 \in [Y]$ de telle sorte que $Y_1 \setminus X_1$ soit dans la classe de Y. Dans ce cas, la règle $X_1 \rightarrow Y_1 \setminus X_1$ va représenter la famille des règles ayant une prémisse dans $[X]$ et un conséquent dans $[Y]$. Dans le pire des cas, c'est-à-dire quand on ne peut pas trouver un motif $X_1 \in [X]$ et un motif $Y_1 \in [Y]$ tels que $Y_1 \setminus X_1$ soient dans la classe de Y, alors dans ce cas et seulement dans ce cas, on se contente de prendre la règle $X_1 \rightarrow Y$ comme représentante (dans la base positive approximative) de la famille des règles que l'on peut générer à partir d'une prémisse dans $[X]$ et d'un conséquent dans $[Y]$. Même si l'intersection de prémisse et de conséquent de cette règle n'est pas vide, elle garde au moins la qualité d'être plus informative que les autres règles (prémisse minimale et conséquent maximal). Nous allons tenir compte de tous ces détails dans la définition de la nouvelle base des règles approximatives selon la mesure M_{GK}.

Proposition 4.7. *Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte d'extraction. Pour tous motifs Z, T de $\mathscr{P}(\mathcal{I})$, Si $Z \rightarrow T \setminus Z$ est M_{GK}-valide au niveau de confiance α, alors il existe deux fermés X et Y tels que :*

$$\begin{cases} \gamma(X) = \gamma(Z), \\ \gamma(Y) = \gamma(T \setminus Z), \\ M_{GK}(X \rightarrow Y) = M_{GK}(Z \rightarrow T \backslash Z), \\ M^{\alpha}_{GK}(X \rightarrow Y) = M^{\alpha}_{GK}(Z \rightarrow T \backslash Z). \end{cases}$$

Preuve.
Soient Z et T deux motifs d'un contexte binaire $\mathcal{K}$. Désignons respectivement par *minSupp* et α le support minimum toléré et le niveau de confiance souhaité. La règle $Z \rightarrow T \setminus Z$ est

une règle valide au niveau de confiance α selon la mesure M_{GK} signifie :

$$\begin{cases} M_{GK}\left(Z \rightarrow T\backslash Z\right) & \geq & M_{GK}^{\alpha}\left(Z \rightarrow T\backslash Z\right), \\ \mathrm{Supp}\left(Z\right) & \geq & minSupp, \\ \mathrm{Supp}\left(T\backslash Z\right) & \geq & minSupp, \end{cases}$$

$$\text{avec } \mathrm{M}_{GK}^{\alpha}\left(Z \rightarrow T\backslash Z\right) = \sqrt{\frac{1}{|\mathcal{O}|}\frac{1-\mathrm{Supp}(Z)}{\mathrm{Supp}(Z)}\frac{\mathrm{Supp}(T \setminus Z)}{1-\mathrm{Supp}(T \setminus Z)}\chi^2_{(1-\alpha)}},$$

et $\chi^2_{(1-\alpha)}$ représente la valeur théorique de χ^2 à 1 degré de liberté et au risque d'erreur $(1-\alpha)$. Comme le support d'un motif est toujours égal au support de sa fermeture, le fait que $T \setminus Z$ soit fréquent nous permet d'affirmer que le support du fermé $\gamma(T \setminus Z)$ dépasse aussi le *minSupp* ($\mathrm{Supp}\left(\gamma\left(T\backslash Z\right)\right) = \mathrm{Supp}\left(T\backslash Z\right) \geq minSupp$). On vient donc de trouver un fermé $Y = \gamma(T \setminus Z)$ tel que $\mathrm{Supp}(Y) \geqslant minSupp$. C'est pareil pour le motif Z, lorsque le support de Z dépasse le *minSupp*, le motif fermé X qui est égal à la fermeture de Z a le même support que Z, donc, $\mathrm{Supp}(\gamma(Z))$ dépasse aussi le *minSupp*. En ce qui concerne la valeur de M_{GK}, nous avons les égalités ci-dessous :

$$\begin{aligned} M_{GK}\left(Z \rightarrow T\backslash Z\right) &= \frac{\mathrm{Conf}\left(Z \rightarrow T\backslash Z\right) - \mathrm{Supp}\left(T\backslash Z\right)}{1-\mathrm{Supp}\left(T\backslash Z\right)} \\ &= \frac{\mathrm{Conf}\left(\gamma\left(Z\right) \rightarrow \gamma\left(T\backslash Z\right)\right) - \mathrm{Supp}\left(\gamma\left(T\backslash Z\right)\right)}{1-\mathrm{Supp}\left(\gamma\left(T\backslash Z\right)\right)} \\ &= \frac{\mathrm{Conf}\left(X \rightarrow Y\right) - \mathrm{Supp}\left(Y\right)}{1-\mathrm{Supp}\left(Y\right)} \\ &= M_{GK}\left(X \rightarrow Y\right). \end{aligned}$$

Et enfin, pour les valeurs critiques, on a :

$$\begin{aligned} \mathrm{M}_{GK}^{\alpha}\left(Z \rightarrow T\backslash Z\right) &= \sqrt{\frac{1}{|\mathcal{O}|}\frac{1-\mathrm{Supp}(Z)}{\mathrm{Supp}(Z)}\frac{\mathrm{Supp}(T \setminus Z)}{1-\mathrm{Supp}(T \setminus Z)}\chi^2_{(1-\alpha)}} \\ &= \sqrt{\frac{1}{|\mathcal{O}|}\frac{1-\mathrm{Supp}(\gamma(Z))}{\mathrm{Supp}(\gamma(Z))}\frac{\mathrm{Supp}(\gamma(T \setminus Z))}{1-\mathrm{Supp}(\gamma(T \setminus Z))}\chi^2_{(1-\alpha)}} \\ &= \sqrt{\frac{1}{|\mathcal{O}|}\frac{1-\mathrm{Supp}(X)}{\mathrm{Supp}(X)}\frac{\mathrm{Supp}(Y)}{1-\mathrm{Supp}(Y)}\chi^2_{(1-\alpha)}} \\ &= \mathrm{M}_{GK}^{\alpha}\left(X \rightarrow Y\right). \end{aligned}$$

Donc, du moment que la règle $Z \rightarrow T\backslash Z$ est valide, on peut toujours trouver deux motifs fermés fréquents X et Y tels que :

$$\begin{cases} M_{GK}\left(Z \rightarrow T\backslash Z\right) & = & M_{GK}\left(X \rightarrow Y\right), \\ M_{GK}^{Cr}\left(Z \rightarrow T\backslash Z\right) & = & M_{GK}^{Cr}\left(X \rightarrow Y\right). \end{cases}$$

On a donc une équivalence entre la validité de $X \rightarrow Y$ et $Z \rightarrow T \setminus Z$. □

Corollaire 4.3 (Axiomes d'inférences pour les règles positives approximatives (PA))**.** *Soient X, Y deux motifs d'un contexte binaire d'extraction $\mathcal{K}$ et α un niveau de confiance fixé. Si la règle $X \rightarrow Y$ est une règle valide au niveau de confiance α ($M_{GK}(X \rightarrow Y) \geqslant \mathrm{M}^{\alpha}_{GK}$, $Supp(X) \geq minSupp$ et $Supp(Y) \geq minSupp$), alors on a :*

$(PA1)$ *Pour tous motifs Z, T tels que : $\gamma(Z) = \gamma(X)$ et $\gamma(T) = \gamma(Y)$, la règle $Z \rightarrow T$ est une règle valide au niveau de confiance α.*

$(PA2)$ *Pour tous motifs Z, T tels que : $\gamma(Z) = \gamma(X)$ et $\gamma(T \setminus Z) = \gamma(Y)$, la règle $Z \rightarrow T \setminus Z$ est une règle valide au niveau de confiance α.*

Remarquons que l'axiome $(PA2)$ est un cas particulier $(PA1)$. On l'utilise lorsque l'on souhaite avoir systématiquement des règles à prémisses et conséquents disjoints.

Proposition 4.8. *Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte binaire d'extraction. Pour tous motifs X et Y, tels que Y est fermé et $\gamma(X) \supseteq Y$, la règle $X \rightarrow Y \setminus X$ est une règle exacte.*

Preuve.
Soient X et Y deux éléments de $\mathscr{P}(\mathcal{I})$ tels que : $\begin{cases} \gamma(Y) = Y, \\ \gamma(X) \supseteq Y. \end{cases}$

$$\begin{aligned} \mathrm{Conf}(X \rightarrow Y \backslash X) &= \frac{\mathrm{Supp}(X \cup Y \backslash X)}{\mathrm{Supp}(X)} \\ &= \frac{\mathrm{Supp}(X \cup Y)}{\mathrm{Supp}(X)} \\ &= \frac{\mathrm{Supp}(\gamma(\gamma(X) \cup Y))}{\mathrm{Supp}(\gamma(X))}. \end{aligned}$$

Comme $\gamma(X) \supseteq Y$, on a : $\gamma(X) \cup Y = \gamma(X)$. Donc, $\mathrm{Conf}(X \rightarrow Y \setminus X) = 1$ et par la suite (selon la proposition 4.3), $\mathrm{M}_{GK}(X \rightarrow Y \setminus X) = 1$. La règle $X \rightarrow Y \setminus X$ est bien une règle exacte. □

Une conséquence immédiate de cette propriété nous permet d'affirmer que pour extraire une règle approximative (de mesure $\mathrm{M}_{GK} \neq 1$), on doit se passer des règles $X \rightarrow Y \setminus X$ avec $\gamma(X) \supseteq Y$.
Avant de voir la nouvelle description de base des règles positives approximatives, nous allons d'abord voir l'ensemble des règles représentantes.

Définition 4.3. *Soient $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte binaire d'extraction et X, Y deux fermés de $\mathscr{P}(\mathcal{I})$. Toutes les règles $r_{ij} : X_i \rightarrow Y_j$ telles que $X_i \in [X]$ et $Y_j \in [Y]$ ont la même mesure et la même valeur critique. Celles qui sont les plus informatives, c'est-à-dire, celles qui ont de prémisse minimale, de conséquent maximal et éventuellement, de prémisse et conséquent disjoints, sont appelées représentantes des règles de prémisse dans $[X]$ et de conséquent dans $[Y]$.*

Exemple 10. *Dans le contexte $\mathcal{K}$ du tableau 2.1, prenons les fermés ABE et $ABEF$. Les éléments de la classe de ABE ceux de $ABEF$ sont :*

$$\begin{aligned} [ABE] &= \{ABE, AE\}, \\ [ABEF] &= \{ABEF, BEF, ABF, AEF, BF, AF, EF\}. \end{aligned}$$

En prenant $X = AE$ (générateur de ABE) et en prenant $Y = ABEF$, le motif $Y \setminus X = BF$ est un élément de $[Y]$. Donc, la règle $AE \rightarrow BF$ est la plus informative (conséquent maximal et de prémisse minimale), elle peut être prise comme représentante de toutes les règles dont la prémisse se trouve dans $[ABE]$ et de conséquent dans $[ABEF]$.
Considérons la classe des fermés ABC et $ABCE$.

$$\begin{aligned} [ABC] &= \{ABC, BC, AC, C\} \\ [ABCE] &= \{ABCE, BCE, ACE, CE\} \end{aligned}$$

En prenant $X = C$ (générateur de ABC), $Y \setminus X = AB$ et $AB \notin [ABCE]$. Dans ce cas, on se contente de la règle $C \rightarrow ABCE$ pour représenter les règles de prémisse dans $[ABC]$ et de conséquent dans $[ABCE]$.

Algorithme de construction d'un ensemble des règles représentantes
Comme nous l'avons fait remarquer plus haut, pour tout motif fermé Y et pour tout générateur G_X d'un fermé X, le motif $Y \setminus G_X$ ne reste pas toujours dans $[Y]$. Pour éviter l'éventualité de perte d'information, il faut à tous prix trouver au moins une représentante des règles dont les prémisses sont dans $[X]$ et le conséquent dans $[Y]$. D'où la necéssité de l'algorithme 2.

Algorithme 2 Recherche des règles représentantes

Entrée : Deux fermés X et Y ($\mathcal{G}_X$: ensemble des générateurs de X)
Sortie : R_{XY} règles représentant celles de prémisse dans $[X]$ et de conséquent dans $[Y]$
1: $R_{XY} = \emptyset$
2: **Pour** Chaque G_X dans $\mathcal{G}_X$ **faire**
3: **Si** $(Y \setminus G_X \in [Y])$ **alors**
4: $R_{XY} = R_{XY} \cup \{G_X \rightarrow Y \setminus G_X\}$
5: **Sinon**
6: $R_{XY} = \{G_X \rightarrow Y\}$
7: **Fin Si**
8: **Fin Pour**

En partant de deux motifs fermés X et Y, l'idée est de trouver des représentantes des règles de prémisse dans $[X]$ et de conséquent dans $[Y]$. L'algorithme commence par parcourir l'ensemble des générateurs de X (ligne 2).
Pour chaque générateur G_X de X, on commence par tester le motif fermé Y (lignes 3 et 4), si $Y \setminus G_X$ reste dans $[Y]$, on prend la règle $G_X \rightarrow Y \setminus G_X$ comme une représentante des règles de prémisse dans $[X]$ et de conséquent dans $[Y]$ obtenue à partir du générateur G_X (ligne 2) et on recommence le processus avec d'autres générateurs de X (sous réserve qu'il en existe encore). Dans le cas contraire (c'est-à-dire, si $Y \setminus G_X \notin [Y]$), on se contente du motif Y et on prend la règle $G_X \rightarrow Y$ comme représentante obtenue à partir du générateur G_X (lignes 6). Remarquons que le nombre des éléments constituant l'ensemble R_{XY} est égale au nombre des générateurs de X. On aurait pu prendre une seule représentante des règles de prémisse dans $[X]$ et de conséquent dans $[Y]$ et dériver les autres règles en utilisant les axiomes d'inférence. Mais, nous avons jugé bon de donner à l'utilisateur toutes les règles valides dont la prémisse est un générateur. En effet, les générateurs ne sont pas comparables ; donc, quand

on a deux règles r_1 et r_2 dont la prémisse de chacune est un générateur d'un fermé X, les deux règles apportent forcément des informations différentes. Il serait intéressant de fournir ces informations avant la dérivation des règles redondantes. Autrement dit, les informations contenues dans la base pourraient être largement suffisantes pour l'interprétation de résultats et par conséquent, le recours à la dérivation des règles redondantes ne sera pratiquement plus nécessaire.
Étant donné qu'un générateur n'est pas nécessairement unique dans sa classe, par rapport à l'ancienne base M_{GK}-valide, établie à partir de prémisse et conséquent fermés, la nouvelle base positive approximative va générer beaucoup plus des règles. D'un autre coté, elle va fournir beaucoup plus d'informations (toutes non redondantes) sans avoir à dériver toutes les règles valides (une opération qui n'est pas très pratique quand on a plusieurs dizaines de variables). Connaissant les règles représentantes pour chaque couple des fermés, on peut définir la nouvelle base positive approximative.

Proposition 4.9 (Nouvelle Base Positive Approximative (NBPA))**.**
Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte binaire d'extraction. Désignons par $FF_{\mathcal{K}}$ l'ensemble des Fermés Fréquents et par $\mathcal{G}_{\mathcal{K}}$ l'ensemble des générateurs des fermés fréquents du contexte $\mathcal{K}$. L'ensemble $NBPA(\alpha)$ défini par :

$$NBPA\,(\alpha) = \{r \in R_{XY} \,/\, X, Y \in FF_{\mathcal{K}},\ X \not\supseteq Y \ \textit{et}\ M_{GK}\,(r) \geq \mathrm{M}^{\alpha}_{GK}\}.$$

où M^{α}_{GK} désigne la valeur critique de M_{GK} de la règle r calculée au niveau de confiance α est une base pour les règles positives approximatives relativement aux axiomes d'inférences $(PA1)$ et $(PA2)$.

Preuve.
Nous allons montrer que l'application des axiomes $(PA1)$ et $(PA2)$ aux règles dans $NBPA$ permet de trouver l'ensemble des règles approximatives valides avec leurs mesures respectives, et qu'aucune règle non valide n'est dérivée de $NBPA$.
Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte binaire d'extraction. Prenons deux motifs X, Y de $\mathscr{P}(\mathcal{I})$ tels que $r : X \to Y$ soit une règle approximative valide. Montrons que si elle n'est pas dans $NBPA$, alors on peut y trouver une règle à partir de laquelle l'application de $(PA1)$ permet de la déduire.

$$r_1 : X \to Y \Leftrightarrow \begin{cases} \mathrm{Supp}\,(X) & \geq & minSupp, \\ \mathrm{Supp}\,(Y) & \geq & minSupp, \\ M_{GK}\,(X \to Y) & \geq & M^{\alpha}_{GK}. \end{cases}$$

Supposons que $r \notin NBPA(\alpha)$.
Puisque X est fréquent ($\mathrm{Supp}\,(X) \geq minSupp$), sa fermeture est aussi un motif fréquent (X et $\gamma(X)$ ont le même support). Donc il existe G_F, générateur du fermé F avec $F = \gamma(X)$ tels que : X soit dans $[F]$ et que G_F soit une partie de X.
De même pour le motif Y, $\gamma(Y)$ est fermé et il contient Y. Comme ces deux motifs (Y et $\gamma(Y)$) ont le même support, le motif $\gamma(Y)$ est donc un motif fréquent. L'application de l'algorithme 2 aux motifs fermés $\gamma(X)$ et $\gamma(Y)$ permet de retrouver une représentante des règles de prémisse dans $[X]$ et de conséquent dans $[Y]$. Selon le contexte étudié, deux cas sont à distinguer.

Premier cas
Lorsqu'il existe un motif $Y_1 \subseteq \gamma(Y)$ tel que $Y_1 \setminus G_F$ soit dans $[Y]$ et par conséquent,

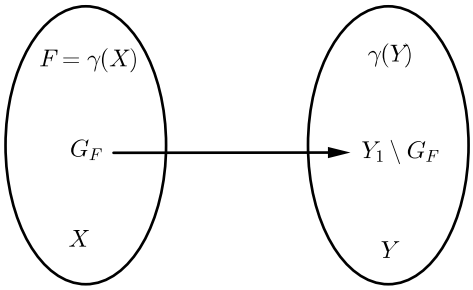


FIGURE 4.4 – Représentante des règles dans $[X]$ *et* $[Y]$

$G_F \rightarrow Y_1 \setminus G_F$ est une représentante des règles entre les deux classes $[X]$ et $[Y]$ (voir Fig. 4.4). Nous avons montré(§ 4.6.1) que pour tous $Z \in [X]$ et $T \in [Y]$, $M_{GK}(X \rightarrow Y) = M_{GK}(Z \rightarrow T)$. De plus, $\text{Supp}(X) = \text{Supp}(Z)$, $\text{Supp}(Y) = \text{Supp}(T)$ et $\text{Conf}(r : X \rightarrow Y) = \text{Conf}(r' : Z \rightarrow T)$, r et r' ont la même valeur critique. D'où les égalités :

$$\begin{cases} M_{GK}(X \rightarrow Y) = M_{GK}(G_F \rightarrow Y_1 \setminus G_F), \\ M^{cr}_{GK}(X \rightarrow Y) = M^{cr}_{GK}(G_F \rightarrow Y_1 \setminus G_F). \end{cases}$$

Comme $X \rightarrow Y$ est une règle approximative valide, la règle représentante $G_F \rightarrow Y_1 \setminus G_F$ l'est aussi et elle est dans $NBPA$. Remarquons maintenant que G_F est un générateur, donc $G_F \subseteq X$ et, $\gamma(Y)$ est fermé, donc $Y_1 \setminus G_F \supseteq \gamma(Y)$; de plus, $\gamma(X) = \gamma(G_F)$ et $\gamma(Y_1 \setminus G_F) = \gamma(Y)$. L'application de $(P21)$ à $G_F \rightarrow Y_1 \setminus G_F$ permet de retrouver la règle $X \rightarrow Y$ ainsi que sa mesure.

Deuxième cas

Dans le cas où $Y \setminus G_F$ n'est pas dans $[Y]$, dans le présent cas, l'algorithme 2 fourni la règle $r : G_F \rightarrow \gamma(Y)$ comme représentante des règles de prémisse dans $[X]$ et de conséquent dans $[Y]$ (voir Fig 4.5).

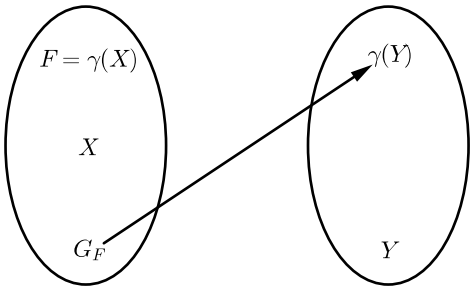


FIGURE 4.5 – Représentante non nécessairement disjointe des règles dans $[X]$ *et* $[Y]$

Puisque $X \supseteq G_F$ et $Y \subseteq \gamma(Y)$, l'application de $(PA1)$ à r permet de retrouver la règle $X \rightarrow Y$ ainsi que sa mesure.

D'un autre coté, soit $X \rightarrow Y$ une règle de $NBPA$ et $Z \rightarrow T$ une règle dérivée de $X \rightarrow Y$ par l'application de $(PA1)$. Selon l'axiome d'inférence (RPA), on a : $\gamma(Z) = \gamma(X)$ et $\gamma(T) = \gamma(Y)$. Ces deux égalités nous permettent d'affirmer, d'une part, que $\text{Supp}(Z) = \text{Supp}(X)$ et $\text{Supp}(T) = \text{Supp}(Y)$. Donc, Z et T sont des motifs fréquents. D'autre part, selon la proposition 4.7, les deux règles $X \rightarrow Y$ et $Z \rightarrow T$ ont la même mesure et valeur critique, donc, $Z \rightarrow T$ est une règle approximative valide.
Étudions maintenant la question de minimalité de $NBPA$. Comme dans le cas de $NBNE$, nous allons définir l'application surjective et non injective R qui associe une règle $r : X \rightarrow Y$ de $NBPA$ à la règle $r' : \gamma(X) \rightarrow \gamma(Y)$ de BNA. Soulignons que l'utilisation de l'application R et le lemme 4.2 permettent aussi de justifier que la base $NBPA$ est bien génératrice. Comme R est non injective, $card(NBPA) \geq card(BPA)$. Prendre une seule règle représentante pour chaque couple des fermées permet d'avoir $NBPA$ même cardinale que BPA. Mais, comme nous l'avons souligné un peu plus haut, la considération de tous les générateurs d'une classe pour la construction des règles représentantes de prémisse dans cette classe est justifiée par le fait que chacune de ces règles représentantes apportent des informations différentes (puisque les générateurs d'une classe ne sont pas comparable entre eux). Avec $NBPA$, l'utilisateur a accès à ces informations sans avoir effectué une quelconque dérivation. □

Exemple 11.
Reprenons le contexte d'extraction décrit dans le tableau 2.1. En prenant un $minSupp = 1/2$*, voici la liste des fermés et des générateurs fréquents :*

$$\begin{aligned} FF_{\mathcal{K}} &= \{ABE, ABC, BE, AB, BD, AD, B, A, D\}, \\ \mathcal{G}_{\mathcal{K}} &= \{AE, C, E, AB, BD, AD, B, A, D\}. \end{aligned}$$

Selon la définition de $NBPA$, pour chaque couple de fermé X et Y, il faut d'abord construire l'ensemble des règles représentantes à partir de l'ensemble des générateurs G_X du motif fermé X et d'un motif Y_1 dans $[Y]$. Ces règles représentantes constituent des candidates à la nouvelle base positive approximative (Tableau 4.6). Pour être validées, les valeurs M_{GK} de ces candidates vont être comparées à leurs valeurs critiques respectives. Les cases croisées contiennent des règles $G_X \rightarrow Y_1$ ($G_X \in \mathcal{G}_X$ et $Y_1 \in [Y]$) qui ne doivent pas se trouver dans $NBPA$ (soit le support de $G_X \rightarrow Y_1$ ne dépasse pas le $minSupp$, soit $\gamma(G_X) \supset Y$, ou encore, les deux motifs constituant la règle sont mutuellement répulsifs, dans ce dernier cas, on doit envisager d'étudier les règles négatives correspondantes).
Nous allons faire une comparaison des règles qui constituent l'ancienne base positive approximative avec celles qui pourraient composer la nouvelle base positive approximative. Dans le tableau 4.7, les règles qui différencient les deux bases (nouvelle et ancienne) sont présentées en caractères rouges. Dans ce simple exemple, on peut faire au moins deux observations. Au niveau des règles candidates, dans l'ancienne comme dans la nouvelle base positive approximative, on a exactement le même nombre de candidats. Par contre, quand c'est possible, prémisse et conséquent de la nouvelle base sont disjoints. À part les quatre premières règles figurant dans le tableau 4.7 à gauche, dans les deux bases, nous avons à chaque fois des règles de même prémisse. Ce constat est juste une particularité du contexte $\mathcal{K}$ du tableau 2.1, mais en réalité, les prémisses et conséquents des règles dans les deux bases sont généralement différents. Dans le présent contexte, plusieurs fermés fréquents sont les uniques éléments composant leurs classes, donc, ils jouent à la fois le rôle de fermé et de générateur. Ceci explique pourquoi on les voit dans la prémisse des règles dans les deux bases. On peut

	ABE	ABC	BE	AB
AE	×	$AE \rightarrow ABC$	×	×
C	$C \rightarrow ABE$	×	×	×
E	$E \rightarrow ABE$	×	×	$E \rightarrow AB$
AB	$AB \rightarrow ABE$	$AB \rightarrow C$	$AB \rightarrow BE$	×
BD	×	×	×	×
AD	×	×	×	×
B	$B \rightarrow AE$	$B \rightarrow AC$	$B \rightarrow E$	$B \rightarrow AB$
A	$A \rightarrow ABE$	$A \rightarrow BC$	×	$A \rightarrow AB$
D	×	×	×	×

	BD	AD	B	A	D
AE	×	×	×	×	×
C	×	×	×	×	×
E	×	×	×	×	×
AB	×	×	×	×	×
BD	×	$BD \rightarrow AD$	×	×	×
AD	$AD \rightarrow BD$	×	×	×	×
B	$B \rightarrow BD$	×	×	×	×
A	×	$A \rightarrow AD$	×	×	×
D	$D \rightarrow BD$	$D \rightarrow AD$	×	×	×

Tableau 4.6 – Génération des candidats à $NBPA$

Ancienne Base	Nouvelle Base	M_{GK}
$BE \rightarrow AB$	$E \rightarrow AB$	0,25
$BE \rightarrow ABE$	$E \rightarrow ABE$	0,50
$ABC \rightarrow ABE$	$C \rightarrow ABE$	0,33
$ABE \rightarrow ABC$	$AE \rightarrow BC$	0,33
$AB \rightarrow ABC$	$AB \rightarrow C$	0,50
$AB \rightarrow BE$	$AB \rightarrow E$	0,25
$B \rightarrow ABE$	$B \rightarrow AE$	0,20
$AB \rightarrow ABE$	$AB \rightarrow ABE$	0,5
$AD \rightarrow BD$	$AD \rightarrow BD$	0,33
$BD \rightarrow AD$	$BD \rightarrow AD$	0,33

Ancienne Base	Nouvelle Base	M_{GK}
$B \rightarrow AB$	$B \rightarrow AB$	0,40
$B \rightarrow BD$	$B \rightarrow BD$	0,20
$A \rightarrow ABE$	$A \rightarrow ABE$	0,20
$A \rightarrow ABC$	$A \rightarrow BC$	0,20
$A \rightarrow AB$	$A \rightarrow AB$	0,40
$A \rightarrow AD$	$A \rightarrow AD$	0,20
$D \rightarrow BD$	$D \rightarrow BD$	0,50
$D \rightarrow AD$	$D \rightarrow AD$	0,50
$B \rightarrow ABC$	$B \rightarrow AC$	0,20
$B \rightarrow BE$	$B \rightarrow E$	0,40

Tableau 4.7 – Comparaison des bases positives approximatives

maintenant soumettre ces règles candidates à leurs valeurs critiques respectives pour pouvoir valider ces liens implicatifs. À titre indicatif et d'exemple, nous allons calculer les valeurs critiques de chacune de ces candidates à $NBPA$. Par ailleurs, remarquons que la valeur critique de M_{GK} est liée au test d'indépendance de χ^2. Or, ce test n'est fiable que lorsque la taille de l'échantillon dépasse la trentaine et que tous les effectifs théoriques sont au moins égaux à cinq. Dans le cas du contexte donné dans le tableau 2.1, la taille de l'échantillon étudié (nombre d'objet ou de transaction) est égale à six ; comme cette quantité est très loin de 30,

les règles qui seront validées par ces valeurs critiques n'auront pas toujours de signification effective, c'est juste un exemple formel.

	Candidates			N_X	N_Y	N_{XY}	M_{GK}	M^{α}_{GK}
r_1	E	$\rightarrow$	AB	4	4	3	0,25	0,42
r_2	AB	$\rightarrow$	ABE	4	3	3	0,5	0,30
r_3	AB	$\rightarrow$	C	4	3	3	0,5	0,30
r_4	AB	$\rightarrow$	E	4	4	3	0,25	0,42
r_5	B	$\rightarrow$	AE	5	3	3	0,2	0,19
r_6	B	$\rightarrow$	AC	5	3	3	0,2	0,19
r_7	B	$\rightarrow$	E	5	4	4	0,4	0,27
r_8	B	$\rightarrow$	AB	5	4	4	0,4	0,27
r_9	B	$\rightarrow$	BD	5	3	3	0,2	0,19
r_{10}	A	$\rightarrow$	ABE	5	3	3	0,2	0,19
r_{11}	A	$\rightarrow$	BC	5	3	3	0,2	0,19
r_{12}	A	$\rightarrow$	AB	5	4	4	0,4	0,27
r_{13}	A	$\rightarrow$	AD	5	3	3	0,2	0,19
r_{14}	D	$\rightarrow$	BD	4	3	3	0,5	0,30
r_{15}	D	$\rightarrow$	AD	4	3	3	0,5	0,30
r_{16}	AD	$\rightarrow$	BD	3	3	2	0,33	0,42
r_{17}	BD	$\rightarrow$	AD	3	3	2	0,33	0,42
r_{18}	C	$\rightarrow$	ABE	3	3	2	0,33	0,42
r_{19}	AE	$\rightarrow$	BC	3	3	2	0,33	0,42
r_{20}	E	$\rightarrow$	ABE	4	3	3	0,5	0,30

Tableau 4.8 – Valeur critique des règles candidates avec $\alpha = 0,7$ (probabilité de ne pas se tromper)

Après avoir choisi une probabilité de ne pas se tromper ($\alpha = 0,7$), on peut voir que dans ce contexte, seules les candidates r_1, r_4 et r_{16} à r_{19} ne sont pas valides (valeur de M_{GK} ne dépasse pas leurs valeurs critiques respectives). Donc, au niveau de confiance α, la $NBPA$ est constituée des 14 règles figurant dans le tableau 4.8.
Généralement, l'utilisateur choisit de prendre une probabilité de ne pas se tromper, c'est-à-dire, le niveau de confiance souhaité dans la prise de décision aux alentours de 95%. Dans le présent exemple, nous avons fait exprès de choisir un niveau de confiance très bas pour avoir beaucoup plus de règles valides dans ce contexte de taille très petite ($Card(\mathcal{O}) = 6$) et ceci, dans le but de mieux comparer l'ancienne et la nouvelle base positive approximative.

4.7 Nouvelle base négative approximative

Comme dans le cas d'étude des règles positives, nous allons apporter quelques modifications à la description de base des règles négatives approximatives M_{GK}-valides. Ces modifications seront basées essentiellement sur les propriétés des motifs fermés, propriétés de la mesure M_{GK}. Soulignons avant tout qu'il existe trois catégories des règles négatives : règles négatives à gauche (de type $\overline{X} \rightarrow Y$), règles négatives à droites (de type $X \rightarrow \overline{Y}$), règles négatives

bilatérales (de type $\overline{X} \rightarrow \overline{Y}$). La relation $\mathrm{M}^{f}_{GK}(X \rightarrow Y) = \mathrm{M}^{f}_{GK}(\overline{Y} \rightarrow \overline{X})$, entre deux motifs X, Y qui se favorisent mutuellement permet de déduire l'ensemble des règles négatives bilatérales à partir des règles positives correspondantes. Donc, aucune autre extraction n'est nécessaire pour avoir les règles négatives bilatérales. Par ailleurs, les propriétés de la mesure M_{GK}, démontrées dans (Totohasina, 2008), permettent d'affirmer que tous les résultats d'étude sur la base des règles négatives à droite sont transposables sur l'extraction de base des règles négatives à droite. De ce fait, nous allons effectuer des études sur la description et l'extraction de la base des règles négatives à droite et ensuite, nous transposerons les résultats pour extraire les règles négatives à gauche. La base des règles négatives est désormais la réunion de ces deux bases, soit :

$$NBNA\,(\alpha) = NBNAG\,(\alpha) \cup NBNAD\,(\alpha)$$

Maintenant, nous allons voir quelques propriétés sur les règles négatives approximatives à droite.

Proposition 4.10. *Soit X, Y deux motifs fermés d'un contexte binaire $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ tels que $P(Y'/X')$ soit plus petit que $P(Y')$ (X défavorise Y). Pour tous motifs Z, T de $\mathscr{P}(\mathcal{I})$ tels que $Z \in [X]$ et $T \in [Y]$, on a :*

$$\begin{cases} Supp(X) &= Supp(Z), \\ Supp(Y) &= Supp(T), \\ \mathrm{M}_{GK}(X \rightarrow \overline{Y}) &= \mathrm{M}_{GK}(Z \rightarrow \overline{T}), \\ \mathrm{M}^{\alpha}_{GK}(X \rightarrow \overline{Y}) &= \mathrm{M}^{\alpha}_{GK}(Z \rightarrow \overline{T}). \end{cases}$$

Preuve. Les deux premières égalités viennent du fait que X et Z, comme Y et T se trouvent dans une même classe (même fermeture).

$$\begin{array}{rcl} \mathrm{Supp}(Z) &=& \mathrm{Supp}(\gamma(Z)) \\ &=& \mathrm{Supp}(X) \\ \mathrm{Supp}(T) &=& \mathrm{Supp}(\gamma(T)) \\ &=& \mathrm{Supp}(Y) \end{array}$$

Prouvons maintenant les deux dernières égalités.

$$\begin{array}{rcl} M_{GK}\left(Z \rightarrow \overline{T}\right) &=& \dfrac{P\left(\overline{T}'/Z'\right) - P\left(\overline{T}'\right)}{1 - P\left(\overline{T}'\right)} \\ &=& \dfrac{1 - P\left(T'/Z'\right) - \left(1 - P\left(T'\right)\right)}{P\left(T'\right)} \\ &=& \dfrac{P\left(T'\right) - P\left(T'/Z'\right)}{P\left(T'\right)} \\ &=& \dfrac{\mathrm{Supp}\left(T\right) - \frac{\mathrm{Supp}(T \cup Z)}{\mathrm{Supp}(Z)}}{\mathrm{Supp}\left(T\right)} \end{array}$$

Or, nous savons que pour tout motif X de $\mathscr{P}(\mathcal{I})$, on a toujours : $Supp(X) = Supp(\gamma(X))$

et, par la suite :

$$\begin{aligned} M_{GK}\left(Z \to \overline{T}\right) &= \frac{\text{Supp}\left(\gamma(T)\right) - \frac{\text{Supp}(\gamma(\gamma(T) \cup \gamma(Z)))}{\text{Supp}(\gamma(Z))}}{\text{Supp}\left(\gamma(T)\right)} \\ &= \frac{\text{Supp}\left(Y\right) - \frac{\text{Supp}(X \cup Y)}{\text{Supp}(X)}}{\text{Supp}\left(Y\right)} \\ &= \frac{P\left(Y'\right) - P\left(Y'/X'\right)}{P\left(Y'\right)} \\ &= \frac{1 - P\left(\overline{Y}'\right) - \left(1 - P\left(\overline{Y}'/X'\right)\right)}{1 - P\left(\overline{Y}'\right)} \\ &= M_{GK}\left(X \to \overline{Y}\right). \end{aligned}$$

De la même manière, avec les valeurs critiques, on a :

$$\begin{aligned} M^{\alpha}_{GK}\left(Z \to \overline{T}\right) &= \sqrt{\frac{1}{|\mathcal{O}|}\frac{1 - \text{Supp}\left(Z\right)}{\text{Supp}\left(Z\right)}\frac{\text{Supp}\left(\overline{T}\right)}{1 - \text{Supp}\left(\overline{T}\right)}\chi^2_{(1-\alpha)}} \\ &= \sqrt{\frac{1}{|\mathcal{O}|}\frac{1 - \text{Supp}\left(Z\right)}{\text{Supp}\left(Z\right)}\frac{1 - \text{Supp}\left(T\right)}{\text{Supp}\left(T\right)}\chi^2_{(1-\alpha)}} \\ &= \sqrt{\frac{1}{|\mathcal{O}|}\frac{1 - \text{Supp}\left(\gamma(Z)\right)}{\text{Supp}\left(\gamma(Z)\right)}\frac{1 - \text{Supp}\left(\gamma(T)\right)}{\text{Supp}\left(\gamma(T)\right)}\chi^2_{(1-\alpha)}} \\ &= \sqrt{\frac{1}{|\mathcal{O}|}\frac{1 - \text{Supp}\left(X\right)}{\text{Supp}\left(X\right)}\frac{1 - \text{Supp}\left(Y\right)}{\text{Supp}\left(Y\right)}\chi^2_{(1-\alpha)}} \\ &= \sqrt{\frac{1}{|\mathcal{O}|}\frac{1 - \text{Supp}\left(X\right)}{\text{Supp}\left(X\right)}\frac{\text{Supp}\left(\overline{Y}\right)}{1 - \text{Supp}\left(\overline{Y}\right)}\chi^2_{(1-\alpha)}} \\ &= M^{\alpha}_{GK}\left(X \to \overline{Y}\right). \end{aligned}$$

□

Proposition 4.11. *Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte binaire d'extraction. Pour tous motifs Z, T de $\mathscr{P}(\mathcal{I})$, si $Z \to \overline{T}$ est valide au niveau de confiance α, alors il existe deux fermés X et Y tels que $\text{M}_{GK}(Z \to \overline{T}) = \text{M}_{GK}(X \to \overline{Y})$.*

Preuve. $r : Z \to \overline{T}$ valide signifie que :

$$\begin{cases} Supp\left(Z\right) \geq minSupp, \\ Supp\left(\overline{T}\right) \geq minSupp, \\ M_{GK}\left(Z \to \overline{T}\right) \geq M^{\alpha}_{GK}\left(Z \to \overline{T}\right). \end{cases}$$

En prenant $X = \gamma(Z)$, le motif X est fermé et fréquent (puisque le motif Z est fréquent et $\text{Supp}(X) = \text{Supp}(\gamma(Z)) = \text{Supp}(Z)$). Le motif $Y = \gamma(T)$ est aussi fermé et selon les propriétés des fermés, on a : $\text{Supp}(\overline{Y}) = \text{Supp}(\overline{\gamma(T)}) = \text{Supp}(\overline{T})$. De plus, selon la proposition 4.11, $\text{M}_{GK}(Z \to \overline{T}) = \text{M}_{GK}(X \to \overline{Y})$ et $\text{M}^{\alpha}_{GK}(Z \to \overline{T}) = \text{M}^{\alpha}_{GK}(X \to \overline{Y})$. Donc, on a une équivalence entre la validité de $Z \to \overline{T}$ et $X \to \overline{Y}$. □

Cette proposition nous amène à l'axiome d'inférence ci-après.

Corollaire 4.4 (Axiome d'inférence pour les règles négatives approximatives (RNA))**.** *Soit X, Y deux motifs d'un contexte binaire d'extraction $\mathcal{K}$ et α un niveau de confiance fixé. Si la règle $X \rightarrow \overline{Y}$ est une règle valide au niveau de confiance α, c-à-d : $M_{GK}\left(X \rightarrow \overline{Y}\right) \geqslant \mathrm{M}^{\alpha}_{GK}$, $Supp\left(X\right) \geq minSupp$ et $Supp\left(\overline{Y}\right) \geq minSupp$), alors on a :*

($NA1$) *Pour tous motifs Z, T tels que : $\gamma\left(Z\right) = \gamma\left(X\right)$ et $\gamma\left(T\right) = \gamma\left(Y\right)$, la règle $Z \rightarrow \overline{T}$ est une règle valide au niveau de confiance α.*

($NA2$) *Pour tous motifs Z, T tels que : $\gamma\left(Z\right) = \gamma\left(X\right)$ et $\gamma\left(T \setminus Z\right) = \gamma\left(Y\right)$, la règle $Z \rightarrow \overline{T \setminus Z}$ est une règle valide au niveau de confiance α.*

Remarquons maintenant que pour deux fermés X et Y, il arrive très souvent qu'ils ne sont pas disjoints. Dans ce cas, si X défavorise Y, alors il est possible que $X \rightarrow \overline{Y}$ soit valide, et l'utilisateur se trouve face à une règle négative valide $X \rightarrow \overline{Y}$, avec $X \cap Y \neq \emptyset$. Pourtant, pour chaque couple de fermés (X, Y) de $\mathscr{P}(\mathcal{I}) \times \mathscr{P}(\mathcal{I})$, nous avons une équivalence entre la validité des règles de prémisses dans $[X]$ et de conséquents dans $[Y]$ (voir fig.4.6), donc au lieu de prendre des règles de type $X \rightarrow \overline{Y}$, avec X, Y fermés (donc maximaux dans leurs classes respectives), on peut choisir d'autres règles, celle qui est plus informative pour représenter toutes les règles de prémisses dans $[X]$ et de conséquents dans $[Y]$.

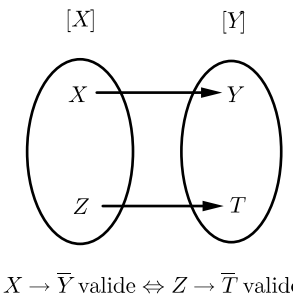


FIGURE 4.6 – Classe des règles négatives

Grâce aux axiomes d'inférence (RNA), il suffit d'avoir une représentante des règles de prémisses dans $[X]$ et de conséquents dans $[Y]$ pour avoir toutes les règles valides de prémisses dans $[X]$ et de conséquents dans $[Y]$. Compte tenu des critiques et remarques faites dans les précédents paragraphes, nous proposons, dans la mesure du possible de choisir des règles plus informatives, c'est-à-dire de prémisses et conséquents minimaux et disjoints. Dans la base des règles négatives approximatives, les prémisses et conséquents seront donc pris dans l'ensemble des générateurs des fermés. Ce choix va engendrer plusieurs conséquences. Il nous permettra, par exemple, d'avoir des règles négatives de prémisses et conséquents minimaux. De plus, avec les générateurs, on minimise considérablement le nombre d'éléments constituant l'intersection des prémisses et conséquents. Autrement dit, avec les générateurs, on a beaucoup plus de chances d'avoir une intersection vide entre une prémisse et un conséquent qu'avec les fermés.

Proposition 4.12. *Soit $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ un contexte binaire d'extraction. Désignons par $\mathcal{G}_X$ et $\mathcal{G}_Y$ l'ensemble des générateurs respectifs des motifs fermés quelconques X et Y et, par $\mathcal{F}$ l'ensemble des fermés du contexte $\mathcal{K}$. La base des règles négatives à droite est :*

$$NBNAD(\alpha) = \left\{ \begin{array}{l} G_X \rightarrow \overline{G_Y} : X, Y \in \mathcal{F}, G_X \in \mathcal{G}_X, G_Y \in \mathcal{G}_Y, \\ G_X \text{ et } \overline{G_Y} \text{ Fréquents et } M_{GK}\left(G_X \rightarrow \overline{G_Y}\right) \geq M^{\alpha}_{GK} \end{array} \right\}.$$

M^{α}_{GK} *désigne la valeur critique de la règle $G_X \rightarrow \overline{G_Y}$ calculé au niveau de confiance α.*

Preuve. Comme dans le cas des règles positives approximatives, nous allons considérer un contexte d'extraction $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ et prenons deux motifs X et Y dans $\mathscr{P}(\mathcal{I})$. Supposons que $X \rightarrow \overline{Y}$ est une règle valide. Montrons que si cette règle n'est pas dans la base des règles négatives approximatives à droite, alors il existe une règle r dans $NBNAD$ telle que l'application de $(NA1)$ ou $(NA2)$ à r permet de retrouver la règle $X \rightarrow \overline{Y}$.

Si la règle $X \rightarrow \overline{Y}$ est valide, alors on a :

$$\left\{ \begin{array}{l} \text{Supp}(X) \geq minSupp, \\ \text{Supp}\left(\overline{Y}\right) \geq minSupp, \\ M_{GK}\left(X \rightarrow \overline{Y}\right) \geq M^{\alpha}_{GK}\left(X \rightarrow \overline{Y}\right). \end{array} \right.$$

Maintenant, considérons deux fermés $Z = \gamma(X)$ et $T = \gamma(Y)$.
Pour tout motif Z_k de la classe $[Z]$ (en l'occurrence le générateur G_Z de $[Z]$), on a toujours : $\text{Supp}(G_Z) = \text{Supp}(Z) = \text{Supp}(X)$. Comme X est fréquent, alors G_Z est aussi un motif fréquent. On sait aussi que $\overline{Y}$ est fréquent et $\text{Supp}(\overline{T}) = \text{Supp}(\overline{Y})$ (puisque $\text{Supp}(T) = \text{Supp}(Y)$). Pour tout T_k dans $[T]$, $\text{Supp}(T_k) = \text{Supp}(T)$, donc $\text{Supp}(\overline{T_k}) = \text{Supp}(\overline{T})$. Si on désigne un générateur de T par G_T (donc, $G_T \in [T]$), la dernière égalité nous permet de déduire que $\text{Supp}(\overline{G_T}) = \text{Supp}(\overline{T})$ et par conséquent, $\text{Supp}(\overline{G_T}) = \text{Supp}(\overline{Y})$. Le motif $\overline{G_T}$ est donc un motif fréquent (voir fig. 4.7).

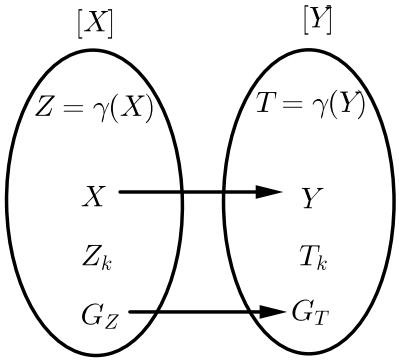


FIGURE 4.7 – Dérivation des règles négatives

Selon la proposition 4.10, les règles $X \rightarrow \overline{Y}$ et $G_Z \rightarrow \overline{G_T}$ ont la même mesure M_{GK} et même valeur critique. Comme X et Y sont fréquents, la règle $r : G_Z \rightarrow \overline{G_T}$ est donc une règle valide. De plus, G_Z et G_T sont des générateurs, donc, la règle r fait partie de la base $NBNAD$. L'application de l'axiome $(NA1)$ à la règle r permet de retrouver la règle $X \rightarrow \overline{Y}$.

D'autre part, toutes les règles dérivées d'une règle valide $G_Z \rightarrow \overline{G_T}$ est une règle valide. En effet, toujours selon la proposition 4.10, $G_Z \rightarrow \overline{G_T}$ et ses dérivées ont la même mesure M_{GK} et même valeur critique et, d'après les propriétés des motifs fermés et la classe des motifs ayant la même fermeture, on a toujours $\mathrm{Supp}(G_Z) = \mathrm{Supp}(X)$ et $\mathrm{Supp}(\overline{G_T}) = \mathrm{Supp}(\overline{Y})$. Donc, on peut conclure que la règle $r' : X \rightarrow \overline{Y}$, dérivée de $r : G_Z \rightarrow \overline{G_T}$ est une règle valide et de plus, r et r' ont les mêmes caractéristiques (M_{GK}, M^{α}_{GK} et Support). On peut donc conclure que la partie $NBNA$ est génératrice de l'ensemble des règles valides au niveau de confiance α. Prendre un seul générateur dans chaque couple de fermé permet d'avoir un ensemble minimal des règles (puisque la famille des classes des fermés d'un contexte forme une partition des motifs de $\mathscr{P}(\mathcal{I})$) relativement aux axiomes $(NA1)$ et $(NA2)$. Mais, comme les générateurs d'un fermé quelconque ne sont jamais comparables, nous avons choisi de mettre dans la base toutes les règles construites à partir de tous les motifs générateurs. □

Exemple 12. *Pour mettre en valeur les différences qu'il pourrait y avoir entre l'ancienne et la nouvelle bases des règles négatives valides selon le couple de mesures Support et* M_{GK}*, nous allons considérer le contexte binaire d'extraction ci-après (cf. Tableau 4.9).*

	A	B	C	D	E	F
o_1	1	1	1	0	1	0
o_2	1	1	1	1	0	0
o_3	0	1	0	1	1	0
o_4	0	1	0	1	1	0
o_5	1	1	1	0	0	1
o_6	1	1	1	0	0	1
o_7	0	1	0	1	1	0
o_8	0	1	0	1	1	0
o_9	0	1	0	1	1	0
o_{10}	0	1	0	1	1	0

Tableau 4.9 – Contexte $\mathcal{K}'$

Prenons deux motifs fermés ABC et BD du contexte décrit dans le tableau 4.9. Examinons la différence entre la nouvelle et l'ancienne base des règles négatives approximatives à droite. Rappelons d'abord que la mesure M_{GK} *possède deux composantes (composante favorisante* M^{f}_{GK} *et défavorisante* M^{d}_{GK}*). On peut montrer que pour tout couple de motifs (X, Y), si X défavorise Y (X favorise $\overline{Y}$), alors $M^{f}_{GK}(X \rightarrow \overline{Y}) = -M^{d}_{GK}(X \rightarrow Y)$. Examinons la règle $ABC \rightarrow \overline{BD}$ (voir fig. 4.8).*

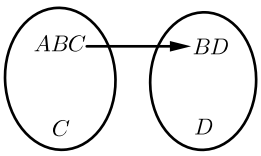


FIGURE 4.8 – Ancienne BNAD

En prenant un $minSupp = 2/5$, les deux motifs ABC et BD sont fréquents. On peut montrer

aussi que ABC défavorise BD, donc, ABC favorise $\overline{BD}$. Par conséquent, on a :

$$\begin{aligned} M_{GK}\left(ABC \to \overline{BD}\right) &= M^{f}_{GK}\left(ABC \to \overline{BD}\right) \\ &= -M^{d}_{GK}\left(ABC \to BD\right) \\ &= -\frac{\frac{n_{ABCD}}{n_{ABC}} - \frac{n_{BD}}{n}}{\frac{n_{BD}}{n}} \\ &= -\frac{\frac{1}{4} - \frac{7}{10}}{\frac{7}{10}} \\ &= 0,64. \end{aligned}$$

Prenons maintenant une précision $\alpha = 0,6$ (risque d'erreur de première espèce $1-\alpha = 0,4$). La valeur théorique de χ^2 est égale à $0,70$, d'où la valeur critique de la mesure M_{GK} de la règle $ABC \to \overline{BD}$:

$$\begin{aligned} \mathrm{M}^{\alpha}_{GK}\left(ABC \to \overline{BD}\right) &= \sqrt{\frac{1}{n}\frac{n - n_{ABC}}{n_{ABC}}\frac{n_{\overline{BD}}}{n - n_{\overline{BD}}}\chi^2_{0,4}} \\ &= \sqrt{\frac{1}{10}\frac{10-4}{4}\frac{3}{10-3}0,7} \\ &= 0,21. \end{aligned}$$

Donc, selon la définition des règles valides au niveau de confiance α (c. f. Définition 4.2 page 74), la règle $ABC \to \overline{BD}$ est une règle négative approximative valide au niveau de confiance $\alpha = 0,6$ ($\mathrm{M}_{GK}(ABC \to \overline{BD}) \geq M^{0,6}_{GK}(ABC \to \overline{BD})$ et de plus, elle fait partie des éléments composant l'ancienne base des règles négatives approximatives au niveau de confiance $\alpha = 0,6$ ($\mathrm{M}_{GK}(ABC \to \overline{BD}) \geq 0,6$). Avec les mêmes fermés ABC et BD, dans la nouvelle base des règles négatives approximatives, on essaye de trouver des liens entre les générateurs (voir fig. 4.9).

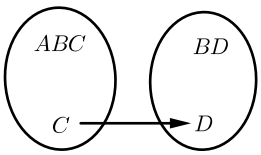


FIGURE 4.9 – Nouvelle BNAD

Puisque C et D font partie des générateurs respectifs de ABC et de BD, mesurons maintenant la règle $C \to \overline{D}$.

$$\begin{aligned} M_{GK}\left(C \to \overline{D}\right) &= M^{f}_{GK}\left(C \to \overline{D}\right) \\ &= -M^{d}_{GK}\left(C \to D\right) \\ &= -\frac{\frac{n_{CD}}{n_C} - \frac{n_D}{n}}{\frac{n_D}{n}} \\ &= -\frac{\frac{1}{4} - \frac{7}{10}}{\frac{7}{10}} \\ &= 0,64. \end{aligned}$$

En ce qui concerne la valeur critique de la mesure de la règle $C \to \overline{D}$, on a :

$$\begin{aligned} \mathrm{M}^{\alpha}_{GK}\left(C \to \overline{D}\right) &= \sqrt{\frac{1}{n}\frac{n-n_C}{n_C}\frac{n_{\overline{D}}}{n-n_{\overline{D}}}\chi^2_{0,4}} \\ &= \sqrt{\frac{1}{10}\frac{10-4}{4}\frac{3}{10-3}0,7} \\ &= 0,21. \end{aligned}$$

D'ailleurs, on a montré dans le paragraphe 4.6.1, qu'entre deux règles de prémisse dans $[X]$ et de conséquent dans $[Y]$, les mesures M_{GK}, ainsi que les valeurs critiques sont égaux. Donc, en prenant un $minSupp = 0,5$ et un seuil de significativité $\alpha = 0,7$, les deux règles $r_1 : ABC \to \overline{BD}$ et $r_2 : C \to \overline{D}$ sont toutes les deux des règles négatives approximatives valides de prémisse dans $[ABC]$ et de conséquent dans $[BD]$. L'application des axiomes d'inférence (RNA) à l'une ou l'autre de ces deux règles permet de retrouver toutes les règles négatives valides de prémisse dans $[ABC]$ et de conséquent dans $[BD]$. Entre les deux règles, nous choisissons de mettre r_2 dans la base. En effet, avec r_2, on a une prémisse et conséquent minimaux et une forte possibilité d'avoir une règle de prémisse et conséquent disjoint, c'est d'ailleurs le cas pour r_2.

4.8 Semi-base M_{GK}-valide

Dans la pratique où l'on analyse plusieurs dizaines ou quelques centaines des variables, si l'interprétation directe de toutes les règles valides constitue une tâche tout simplement irréalisable, les bases des règles, telles qu'elles sont définies actuellement, offrent déjà une bonne synthèse tout à fait exploitable représentant toutes les informations émanant de l'ensemble des règles valides. Néanmoins, toujours dans le but de pouvoir réduire la taille des règles à interpréter, nous proposons dans cette partie, une réduction possible des bases que nous venons de définir. Cette réduction est fondée sur une relation d'ordre que nous allons définir dans l'ensemble des règles constituant une base. En effet, si l'utilisateur se fixe une marge de tolérance (un *minSupp* et une probabilité α de ne pas se tromper), c'est qu'il est prêt à prendre une décision relativement à ces marges de tolérance.
Prenons l'exemple des deux règles valides $r_1 : X \to Y$ et $r_2 : Z \to Y$ avec $Z \supset X$. Supposons que $\mathrm{M}_{GK}(r_1) = \alpha_1$ et $\mathrm{M}_{GK}(r_2) = \alpha_2$ avec $\alpha_1 \neq \alpha_2$. Puisque tous les deux sont valides, on a : $\alpha_1 \geq \mathrm{M}^{\alpha}_{GK}(r_1)$ et $\alpha_2 \geq \mathrm{M}^{\alpha}_{GK}(r_2)$. Certes, les deux règles ont des degrés de précision différents ($\alpha_1 \neq \alpha_2$), mais si l'utilisateur ne se réfère qu'au niveau de précision qu'il a fixé, c'est-à-dire qu'il est prêt à prendre en compte et sans autre distinction (sur le même pied) toutes les informations valides relativement à ce niveau de confiance, la règle r_2, quelle que soit sa mesure, n'apporte aucune information supplémentaire si on la compare à la règle r_1. La règle r_2 est donc devenue une règle redondante à l'égard de r_1 parce qu'elle n'apporte que des informations moins générales que r_1. C'est ce type de règle que nous allons écarter de la base, pour pouvoir constituer ce que nous appelons semi-base M_{GK}-valide. Le terme semi-base vient du fait qu'une fois ces règles redondantes seront écartées, les règles restantes ne constituent pas tout à fait une base dans le sens où certaines règles valides ne pourront plus en être déduites ; cependant ces règles sont toutes redondantes si on ne se réfère qu'à la précision souhaitée par l'utilisateur.

Avant de détailler la construction des semi-bases M_{GK}-valides, nous allons d'abord définir une relation d'ordre dans les bases des règles M_{GK}-valides.

Relation d'ordre dans l'ensemble des règles positives exactes

Définition 4.4.
Soit $\mathcal{E}$ un ensemble des règles positives exactes d'un certain contexte binaire d'extraction $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$. Soient, X_1, X_2, Y_1 et Y_2 des parties de $\mathcal{I}$ telles que les règles $r_1 : X_1 \to Y_1$ et $r_2 : X_2 \to Y_2$ soient dans $\mathcal{E}$. On dit que r_1 est dominée par r_2, notée par $r_1 \prec r_2$ si r_1 et r_2 sont valides $(\min(Supp(r_1), Supp(r_2)) \geq minSupp)$, *$X_2 \subseteq X_1$ et $Y_2 \supseteq Y_1$.*

Plus explicitement, en exploitant la distributivité entre les connecteurs logiques « ou, et » la définition 4.4 peut s'énoncer comme suit :

$$\begin{cases} r_1, r_2 \in \mathcal{E} \\ r_1 : X_1 \to Y_1 \prec r_2 : X_2 \to Y_2 \end{cases} \Leftrightarrow \begin{cases} r_1 \text{ et } r_2 \text{ sont valides,} \\ X_2 \subset X_1 \text{ et } Y_2 = Y_1 \text{ ou} \\ X_2 \subset X_1 \text{ et } Y_2 \supset Y_1 \text{ ou} \\ X_2 = X_1 \text{ et } Y_2 \supset Y_1 \text{ ou} \\ X_2 = X_1 \text{ et } Y_2 = Y_1. \end{cases}$$

Le dernier est un cas trivial auquel r_1 et r_2 représentent une même règle.

Proposition 4.13. *L'ensemble $\mathcal{E}$ muni de la relation $\prec$ est un ensemble partiellement ordonné.*

Preuve.
Montrons que $\prec$ est une relation d'ordre.

Réflexivité Par sa construction, la relation $\prec$ est réflexive. En effet, pour toute règle $r : X \to Y$ dans $\mathcal{E}$, on peut écrire $X \subseteq X$ et $Y \supseteq Y$. Si r est valide, on a : $r \prec r$.

Transitivité Soient $r_1 : X_1 \to Y_1, r_2 : X_2 \to Y_2$ et $r_3 : X_3 \to Y_3$ trois règles valides de $\mathcal{E}$. Supposons que $r_1 \prec r_2$ et $r_2 \prec r_3$, montrons que $r_1 \prec r_3$.

$$r_1 \prec r_2 \Leftrightarrow X_1 \supseteq X_2 \text{ et } Y_1 \subseteq Y_2 \tag{4.6}$$

De même pour r_2 et r_3, on a :

$$r_2 \prec r_3 \Leftrightarrow X_2 \supseteq X_3 \text{ et } Y_2 \subseteq Y_3. \tag{4.7}$$

En combinant 4.6 et 4.7, on a :

$$\begin{cases} X_1 \supseteq X_2 \supseteq X_3, \\ Y_1 \subseteq Y_2 \subseteq Y_3. \end{cases} \tag{4.8}$$

Selon les relations 4.8 et en se servant de la transitivité de la relation d'inclusion, on a : $X_1 \supseteq X_3$ et $Y_1 \subseteq Y_3$. Comme r_1 et r_3 sont valides :

$$\begin{cases} r_1, r_3 \text{ sont valides} \\ X_1 \supseteq X_3 \text{ et } Y_1 \subseteq Y_3 \end{cases} \Leftrightarrow r_1 \prec r_3$$

D'où la transitivité de la relation « dominée par $\prec$ ».

Vérifions que la relation $\prec$ est antisymétrique
Soit r_1 et r_2 deux règles valides de $\mathcal{E}$.
Supposons que $r_1 \prec r_2$ et $r_2 \prec r_1$.

$$r_1 \prec r_2 \Leftrightarrow \begin{cases} X_1 \supseteq X_2 \\ Y_1 \subseteq Y_2 \end{cases} \tag{4.9}$$

$$r_2 \prec r_1 \Leftrightarrow \begin{cases} X_2 \supseteq X_1 \\ Y_2 \subseteq Y_1 \end{cases} \tag{4.10}$$

En combinant 4.9 et 4.10, on a :

$$\begin{cases} X_1 \supseteq X_2 \text{ et } X_2 \supseteq X_1, \\ Y_1 \subseteq Y_2 \quad \text{et } Y_2 \subseteq Y_1. \end{cases}$$

Comme la relation d'inclusion est antisymétrique, on a : $X_1 = X_2$ et $Y_1 = Y_2$.
Donc, si $r_1 \prec r_2$ et $r_2 \prec r_1$, alors $r_1 \equiv r_2$ (r_1 et r_2 représentent une même règle).
La relation $\prec$ est à la fois réflexive, transitive et antisymétrique, donc elle constitue une relation d'ordre et l'ensemble $\mathcal{E}$ muni de la relation $\prec$ est un ensemble ordonné. Comme les prémisses ou les conséquents des règles dans $\mathcal{E}$ ne sont pas tous comparables, la relation d'ordre $\prec$ dans $\mathcal{E}$ est une relation d'ordre partiel. □

Proposition 4.14 (Relation « dominée par » dans l'ensemble des règles dérivées)**.**

1. *Soit $r : X \to Y$ une règle exacte valide. Pour toute règle $dr : Z \to T$, dérivée de la règle r par l'application de l'axiome d'inférence (RPE), la règle dr est toujours dominée par la règle r $(dr \prec r)$.*
2. *Désignons par $\mathcal{B}$ la nouvelle base positive exacte. Soit $r_1 : X_1 \to Y_1$ et $r_2 : X_2 \to Y_2$ deux éléments de $\mathcal{B}$. Si r_1 est dominée par r_2 $(r_1 \prec r_2)$, alors pour toute règle $dr_1 : Z \to T$, dérivée de la règle r_1, dr_1 est toujours dominée par r_2 $(dr_1 \prec r_2)$.*

Preuve.
Rappelons qu'une règle $r : X \to Y$ est exacte et valide selon M_{GK} signifie que :

$$\begin{cases} \mathrm{Supp}\,(X) & \geq & minSup, \\ \mathrm{Supp}\,(Y) & \geq & minSup, \\ M_{GK}\,(X \to Y) & = & 1. \end{cases}$$

$dr : Z \to T$ est dérivée d'une règle exacte $r : X \to Y$ signifie que Z et T sont dans la classe de X([X] qui est aussi la classe de Y d'ailleurs). Comme X est un générateur et Y est fermé (le plus grand motif de la classe), on a toujours : $Z \supseteq X$ et $T \subseteq Y$, et, selon la définition de la relation $\prec$, dr est toujours dominée par r $(dr \prec r)$.
Considérons maintenant deux règles $r_1 : X_1 \to Y_1$ et $r_2 : X_2 \to Y_2$ d'une base positive exacte $\mathcal{B}$. Supposons que r_1 est dominée par r_2 et prenons une règle dr_1, dérivée de la règle r_1 par l'application de l'axiome (RPE). Montrons que dr_1 est toujours dominée par r_2.

$$\begin{cases} r_1, r_2 \in \mathcal{B} \\ r_1 \prec r_2 \end{cases} \Leftrightarrow \begin{cases} M_{GK}\,(r_1) = M_{GK}\,(r_2) = 1 \\ X_1 \supseteq X_2 \text{ et } Y_1 \subseteq Y_2 \end{cases} \tag{4.11}$$

Selon le premier point de la proposition 4.14, si a règle dr_1 est la dérivée d'une certaine règle r_1 par l'application de l'axiome (RPE), alors on a toujours : $dr_1 \prec r_1$, qui peut encore se traduire par l'équivalence :

$$dr_1 : Z \to T \prec r_1 : X_1 \to Y_1 \Leftrightarrow \begin{cases} M_{GK}(dr_1) = M_{GK}(r_1) = 1 \\ Z \supseteq X_1 \\ T \subseteq Y_1. \end{cases} \tag{4.12}$$

En combinant (4.11) et (4.12), nous avons : $\begin{cases} Z \supseteq X_1 \text{ et } X_1 \supseteq X_2, \\ T \subseteq Y_1 \text{ et } Y_1 \subseteq Y_2, \\ M_{GK}(dr_1) = M_{GK}(r_1) = M_{GK}(r_2) = 1. \end{cases}$

Ces relations nous permettent d'écrire : $\begin{cases} Z \supseteq X_2, \\ T \subseteq Y_2, \\ M_{GK}(dr_1) = M_{GK}(r_2). \end{cases}$

Comme dr_1 est une dérivée de r_1 par l'application de l'axiome (RPE), $dr_1 : Z \to T$ est une règle exacte valide, de plus $Z \supseteq X_2$ et $T \subseteq Y_2$. Donc, toute règle dérivée dr_1 de r_1 est dominée par r_2 $(dr_1 \prec r_2)$. □

Relation d'ordre dans l'ensemble des règles positives approximatives

Avant de voir les propriétés connexes à la relation d'ordre permettant d'ordonner les règles approximatives, nous allons comparer les avantages et inconvénients de l'utilisation des axiomes d'inférence $(PA1)$ et $(PA2)$ pour dériver les règles approximatives.
Nous avons vu au paragraphe 4.3 qu'il existe deux manières de dériver une règle approximative :

D1 : À partir d'une règle représentante $r : X \to Y$, on se met dans la classe de X et dans celle de Y pour trouver les règles dérivées. Selon $(PA1)$, la règle $dr : Z \to T$ avec $Z \in [X]$ et $T \in [Y]$ est une règle dérivée de r. Cette méthode de dérivation a l'avantage d'être exhaustive. C'est à dire qu'avec $(PA1)$, on peut dériver toutes les règles possibles et imaginables de prémisse dans $[X]$ et de conséquent dans $[Y]$, y compris les règles de prémisse et conséquent non disjoints. La difficulté qu'on peut rencontrer dans l'utilisation de $(PA1)$ réside sur le fait que la relation « dominée par » telle qu'elle est définie dans la description des semi-bases positives exactes ne sera plus stable dans l'ensemble des règles approximatives dérivées (la règle représentante ne domine pas forcément ses règles dérivées). En effet, le conséquent d'une règle représentante n'est pas forcément le plus grand motif (au sens de l'inclusion) dans sa classe. Par la suite, le conséquent d'une règle dérivée peut être un sous-ensemble comme il peut être un sur-ensemble du conséquent de la règle représentante.

D2 : Comme dans le cas de D1, à partir d'une règle représentante $X \to Y$, on se met dans $[X]$ et $[Y]$ pour trouver les règles dérivées. Selon $(PA2)$, les règles dérivées sont de type $dr : Z \to T \setminus Z$ avec $Z \in [X]$ et $T \setminus Z \in [Y]$. L'inconvénient ou l'avantage (selon les points de vues) de cette méthode de dérivation réside sur le fait qu'on ne peut dériver que des règles à prémisse et conséquent disjoints. De plus, nous allons voir que les conséquents des règles dérivées sont des sous-ensembles du conséquent de la règle représentante et que, les prémisses des règles dérivées sont des sur-ensembles de prémisse d'une règle représentante.

Ainsi, dans la suite de cette étude, nous avons choisi d'utiliser la deuxième méthode de dérivation afin d'avoir des règles de prémisse et conséquent disjoints et la possibilité de définir une relation d'ordre dans l'ensemble des règles dérivées.

Définition 4.5. *Soit $\mathcal{A}$ un ensemble des règles positives approximatives d'un certain contexte binaire d'extraction $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$. Soient X_1, X_2, Y_1 et Y_2 des parties de $\mathcal{I}$ telles que $r_1 : X_1 \rightarrow Y_1$ et $r_2 : X_2 \rightarrow Y_2$ soient dans $\mathcal{A}$. On dit que r_1 est dominée par r_2 au niveau de confiance α et que l'on note par $r_1 \prec_\alpha r_2$ si :*

$$\begin{cases} M_{GK}(r_1) \geq M^{\alpha}_{GK}, \\ M_{GK}(r_2) \geq M^{\alpha}_{GK}, \\ X_1 \supseteq X_2 \text{ et } Y_1 \subseteq Y_2 . \end{cases}$$

En exploitant les propriétés de la relation d'inclusion ensembliste comme dans le cas des règles exactes, on peut montrer que l'ensemble $\mathcal{A}$, muni de la relation « dominée par » est un ensemble ordonné.

Proposition 4.15 (Relation « dominée par » dans l'ensemble des règles dérivées)**.**
Désignons par $\mathcal{A}$ la nouvelle base positive approximative d'un certain contexte d'extraction $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$.

- *Soit $r : X \rightarrow Y \setminus X$ un élément de $\mathcal{A}$. Pour toute règle $dr : Z \rightarrow T \setminus Z$, dérivée de la règle r par l'application de l'axiome d'inférence (RPA), la règle dr est toujours dominée par la règle r ($dr \prec_\alpha r$).*
- *Soit $r_1 : X_1 \rightarrow Y_1$ et $r_2 : X_2 \rightarrow Y_2$ deux éléments de $\mathcal{A}$. Si r_1 est dominée par r_2 ($r_1 \prec_\alpha r_2$), alors pour toute règle $dr_1 : Z \rightarrow T \setminus Z$, dérivée de la règle r_1, dr_1 est toujours dominée par r_2 ($dr_1 \prec_\alpha r_2$).*

Preuve. Soit $r : X \rightarrow Y \setminus X$ une règle de $\mathcal{A}$, deux cas sont à envisager.

1er Cas : $\begin{cases} Y \in FF_{\mathcal{K}}, \ \ X \in \mathcal{G}_{\gamma(X)}, \\ M_{GK}(X \rightarrow Y \backslash X) \geq \mathrm{M}^{\alpha}_{GK}. \end{cases}$

$\mathcal{G}_{\gamma(X)}$ désigne l'ensemble des générateurs du motif fermé $\gamma(X)$.
Ce cas est réalisé lorsque $Y \setminus X \in [Y]$ (c'est-à-dire $\gamma(Y \setminus X) = \gamma(Y)$). Soit $dr : Z \rightarrow T \setminus Z$ une règle dérivée de r par l'application de l'axiome d'inférence (RPA).
Dans ce cas, $\gamma(Z) = \gamma(X)$ et $\gamma(T \setminus Z) = \gamma(Y \setminus X) = \gamma(Y)$. Comme X est un générateur, Z ne peut être qu'un sur-ensemble de X.

$$Z \supseteq X \tag{4.13}$$

Comme Y est un motif fermé, donc il est maximal dans sa classe, T ne peut être qu'un sous-ensemble de Y. À partir des inclusion $T \subseteq Y$ et $X \subseteq Z$, on obtient :

$$T \backslash Z \subseteq Y \backslash X. \tag{4.14}$$

De plus, selon la proposition 4.7 :

$$\begin{cases} M_{GK}(r) = M_{GK}(dr), \\ M^{\alpha}_{GK}(r) = M^{\alpha}_{GK}(dr). \end{cases} \tag{4.15}$$

En combinant les relations (4.13), (4.14) et (4.15), on arrive à conclure que $dr \prec_\alpha r$. Donc, toute règle dr, dérivée de la règle r par l'application de l'axiome d'inférence $(PA2)$ est

dominée par r.

2e Cas : $\begin{cases} Y \notin FF_{\mathcal{K}},\ \ X \in \mathcal{G}_{\gamma(X)}, \gamma(Y) \setminus X \notin [Y], \\ Y \subset \gamma(Y) \text{ tel que } Y \setminus X \in [Y] \text{ et } \nexists Y_1 \supset Y \text{ tel que } Y_1 \setminus X \in [Y], \\ M_{GK}\left(X \rightarrow Y \backslash X\right) \geq M^{\alpha}_{GK}. \end{cases}$

Dans ce cas, pour toutes règles dérivées $dr : Z \rightarrow T \setminus Z$:

1. $Z \supseteq X$, car $\begin{cases} X \text{ est un générateur} \\ Z \text{ est obtenu par l'application de } (PA2), \text{ donc } Z \in [X] \end{cases}$
2. $T \setminus Z \in [Y]$, comme il n'existe pas un motif Y_1 tel que $Y_1 \setminus X$ soit dans $[Y]$, $T \setminus Z$ ne peut être qu'un sous ensemble de $Y \setminus X$.

De plus :

$$\begin{cases} Z \in [X] \\ T \backslash Z \in [Y] \end{cases} \Leftrightarrow \begin{cases} \gamma\left(Z\right) = \gamma\left(X\right) \\ \gamma\left(Z \backslash T\right) = \gamma\left(Y\right) \end{cases}$$

D'après cette équivalence et selon la proposition 4.7 :

$$\begin{cases} M_{GK}\left(r\right) = M_{GK}\left(dr\right), \\ M^{\alpha}_{GK}\left(r\right) = M^{\alpha}_{GK}\left(dr.\right) \end{cases}$$

Autrement dit, $Z \supseteq X$, $T \setminus Z \subseteq Y \setminus X$ et, r et dr ont la même mesure et même valeur critique, donc $dr : Z \rightarrow T \setminus Z \prec_{\alpha} r : X \rightarrow Y \setminus X$.

Prenons maintenant le cas d'une règle $r : X \rightarrow Y$ de $\mathcal{A}$ avec $X \cap Y \neq \emptyset$. On peut avoir ce type de règle lorsqu'il n'existe pas un $Y_1 \subseteq \gamma(Y)$ tel que $Y_1 \setminus X$ soit dans $[Y]$. Dans cette situation, selon la description dans l'algorithme 2, on prend un générateur G_X de X et le fermé Y pour construire une règle représentante de toutes celles de prémisse dans $[X]$ et de conséquent dans $[Y]$. Donc, si $r : X \rightarrow Y$ (avec $X \cap Y \neq \emptyset$) est dans la base $\mathcal{A}$, c'est que X un générateur et Y un fermé. Dans ce cas, les règles $dr : Z \rightarrow T$, dérivées de r par l'application de l'axiome $(PA1)$ vérifient les égalités : $\gamma(Z) = \gamma(X)$ et $\gamma(T) = \gamma(Y)$. Comme X est un générateur et Y un fermé, on a toujours :

$$\begin{cases} Z \supseteq X, \\ T \subseteq Y. \end{cases}$$

Comme les mesures et les valeurs critiques sont les mêmes, selon la définition de la relation « dominée par », on a : $dr \prec_{\alpha} r$.

Dans ces trois cas, les règles dérivées sont toutes dominées par la règle représentante (celle qui est dans la base).

Le second point de la proposition est une conséquence immédiate de celui du premier. En effet, si on prend deux règles r_1 et r_2 dans $\mathcal{A}$ telles que $r_1 \prec_{\alpha} r_2$, selon le premier point, r_1 domine toutes ses règles dérivées. Or, on a montré que la relation « dominée par » est une relation d'ordre, donc elle est transitive. En exploitant cette propriété de transitivité, on a : $dr_1 \prec_{\alpha} r_1$ et $r_1 \prec_{\alpha} r_2$, donc $dr_1 \prec_{\alpha} r_2$. Nous allons maintenant nous servir des propriétés de relation d'ordre « dominée par » pour definir les semi-bases M_{GK}-valides. □

Relation d'ordre dans l'ensemble des règles négatives approximatives

Comme dans l'ensemble des règles positives, dans ce paragraphe, nous cherchons à comparer les règles négatives approximatives valides entre elles. En effet, toujours dans le souci d'avoir

un nombre des règles valides trop élevé, nous allons chercher des moyens pour réduire le nombre des règles présentées à l'utilisateur et tout cela doit se passer sans aucune perte d'information. L'idée est de comparer deux règles selon leurs prémisses et leurs conséquents. Prenons l'exemple des deux règles négatives approximatives $r_1 : AB \rightarrow \overline{C}$ et $r_2 : A \rightarrow \overline{C}$. Selon r_1, la présence simultanée des motifs A et B constitue une condition suffisante pour que le motif négatif $\overline{C}$ soit présent dans les transactions (autrement dit, le motif C n'est pas dans les transactions). D'autre part, selon r_2, la présence de A suffit pour déduire l'absence de C. Par rapport à ces deux informations, si on souhaite avoir une condition suffisante pour assurer l'absence de C, l'information apportée par la règle r_2 suffit. Donc, pour les règles négatives à droites, les prémisses sont des motifs positifs. Dans le même ordre d'idées que pour les règles positives, plus le cardinal des éléments constituant la prémisse diminue, plus l'informativité de la règle augmente.
Du côté des motifs conséquents, nous avons vu au paragraphe 4.3.3 qu'une règle négative à droite apporte de plus en plus d'informations au fur et à mesure de la diminution (en terme de cardinalité) des éléments constituant le motif conséquent. Ainsi, nous pouvons définir une seule relation d'ordre dans l'ensemble des règles négatives à droites.

Définition 4.6.
Soit $\mathcal{N}$ un ensemble des règles négatives approximative d'un certain contexte d'extraction $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$. Soient, X_1, X_2, Y_1 et Y_2 des parties de $\mathcal{I}$ telles que les règles $r_1 : X_1 \rightarrow \overline{Y_1}$ et $r_2 : X_2 \rightarrow \overline{Y_2}$ soient dans $\mathcal{N}$. On dit que r_1 est dominée par r_2 au niveau de confiance α, notée par $r_1 \prec_\alpha r_2$ lorsque r_1 et r_2 sont valides au niveau de confiance α ($\mathrm{M}_{GK}(r_1) \geq \mathrm{M}^{\alpha}_{GK}(r_1)$ et $\mathrm{M}_{GK}(r_2) \geq \mathrm{M}^{\alpha}_{GK}(r_2)$) et de plus, $X_2 \subseteq X_1$ et $Y_2 \subseteq Y_1$.

En exploitant la distributivité entre les connecteurs logiques « ou, et », la définition 4.6 peut être énoncée comme suit :

$$\begin{cases} r_1, r_2 \in \mathcal{N} \\ r_1 : X_1 \rightarrow \overline{Y_1} \prec_\alpha r_2 : X_2 \rightarrow \overline{Y_2} \end{cases} \Leftrightarrow \begin{cases} r_1 \text{ et } r_2 \text{ sont valides ,} \\ X_2 \subset X_1 \text{ et } Y_2 = Y_1 \text{ ou} \\ X_2 \subset X_1 \text{ et } Y_2 \subset Y_1 \text{ ou} \\ X_2 = X_1 \text{ et } Y_2 \subset Y_1 \text{ ou} \\ X_2 = X_1 \text{ et } Y_2 = Y_1. \end{cases}$$

Le dernier est un cas trivial auquel r_1 et r_2 représentent une même règle. La définition 4.6 nous sera utile dans la description des semi-bases des règles négatives.

Proposition 4.16. *La relation « dominé par ($\prec_\alpha$) » définie dans l'ensemble des règles négatives approximative est une relation d'ordre partiel.*

Preuve. En utilisant les propriétés de l'inclusion, nous pouvons montrer que la relation $\prec_\alpha$ est bien une relation d'ordre. En effet, elle est réflexive puisque selon la définition 4.6, pour toute règle $r : X \rightarrow \overline{Y}$ de l'ensemble des règles négatives, r est en relation avec r ($X \subseteq X$ et $Y \subseteq Y$). Prenons maintenant trois règles $r_1 : X_1 \rightarrow \overline{Y_1}, r_2 : X_2 \rightarrow \overline{Y_2}$ et $r_3 : X_3 \rightarrow \overline{Y_3}$ telles que $r_1 \prec_\alpha r_2$ et $r_2 \prec_\alpha r_3$.

$$\begin{cases} r_1 \prec_\alpha r_2 \\ r_2 \prec_\alpha r_3 \end{cases} \Leftrightarrow \begin{cases} r_1,\ r_2 \text{ et } r_3 \text{ sont valides} \\ X_2 \subseteq X_1 \text{ et } Y_2 \subseteq Y_1 \\ X_3 \subseteq X_2 \text{ et } Y_3 \subseteq Y_2 \end{cases}$$

En utilisant la transitivité de la relation d'inclusion, nous avons : $X_3 \subseteq X_1$ et $Y_3 \subseteq Y_1$, donc r_1 est dominé par r_3. Enfin, prenons deux règles négatives $r_1 : X_1 \rightarrow \overline{Y_1}$ et $r_2 : X_2 \rightarrow \overline{Y_2}$ telles que $r_1 \prec_\alpha r_2$ et $r_2 \prec_\alpha r_1$.

$$r_1 \prec_\alpha r_2 \Leftrightarrow \begin{cases} r_1 \, et \, r_2 \;\; sont \; valides \\ X_2 \subseteq X_1 \; et \; Y_2 \subseteq Y_1 \end{cases} \tag{4.16}$$

$$r_2 \prec_\alpha r_1 \Leftrightarrow \begin{cases} r_1 \, et \, r_2 \;\; sont \; valides \\ X_1 \subseteq X_2 \; et \; Y_1 \subseteq Y_2 \end{cases} \tag{4.17}$$

En combinant (4.16) et (4.17), nous pouvons déduire que X_1 ne peut être que X_2 et Y_1 ne peut être que Y_2. Donc, r_1 et r_2 désignent une même règle. La relation « dominé par » est donc antisymétrique. La relation $\prec_\alpha$ est donc une relation d'ordre. □

Relation d'ordre dans l'ensemble des règles négatives dérivées
Selon la définition 4.12, une règle $r : X \rightarrow \overline{Y}$ dans $NBNA$ signifie que X et Y sont des générateurs. Selon les deux axiomes $(NA1)$ et $(NA2)$, une règle $dr : Z \rightarrow \overline{T}$, dérivée de r est une règle de prémisse dans $[X]$ et de conséquent dans $[Y]$. Il en résulte que :

1. soit X et Z sont comparables, et nous avons nécessairement $X \subseteq Z$,
2. soit X et Z ne sont pas comparables (il peut y avoir plusieurs générateurs dans une classe), dans ce cas, il existe d'autres générateurs dans $[X]$ qui sont comparables à Z.

On fait les mêmes constats pour les motifs conséquents. Donc, dans tous les cas, les dérivées sont dominées par les règles qui sont dans la $NBNA$.

4.8.1 Semi-base positive exacte

Compte tenu de la relation d'ordre « dominée par » définie dans la base des règles positives exactes, il est possible d'ordonner, même partiellement les éléments de $NBPE$. Or, selon la définition 3.8 d'une règle redondante établie au chapitre 3, entre deux règles de même mesure et de prémisse et conséquent comparables, la plus informative est celle qui a une prémisse minimale et un conséquent maximal. Donc, une règle r_1, dominée par une autre règle r_2 est forcément moins informative que la règle dominante. Autrement dit, si $r_1 : X_1 \rightarrow Y_1 \prec_\alpha r_2 : X_2 \rightarrow Y_2$, alors r_2 est plus informative que r_1, puisque :

$$\begin{cases} M_{GK}(r_1) = M_{GK}(r_2), \\ X_1 \supseteq X_2 \; et \; Y_1 \subseteq Y_2. \end{cases}$$

Par rapport à ce constat, nous proposons dans Semi-Base Positive Exacte $(SBPE)$ de ne garder que les règles dominantes. D'où la définition ci-dessous :

$$SBPE = \{r : G \rightarrow F \backslash G, \, telle \, que \, F \in FF_{\mathcal{K}}, \; G \in \mathcal{G}_F \; et \; \nexists r' / \; r \prec_\alpha r'\}.$$

Exemple 13.
Reprenons le contexte $\mathcal{K}$ donné dans le tableau 2.1. En prenant un minSupp = 1/3, la nouvelle base positive exacte est donné dans le tableau 4.4 (droite). Examinons de près les règles $r_1 : E \rightarrow B, r_2 : AE \rightarrow B, r_3 : DE \rightarrow B$ et les deux règles $r_3 : C \rightarrow AB, r_4 : CE \rightarrow AB$. Ces règles sont toutes exactes. Entre r_1, r_2 et r_3, les prémisses sont comparables, et

c'est la règle r_1 qui a une prémisse minimale, même constat entre r_4 et r_5. Si on utilise la relation d'ordre « dominée par », on peut écrire $r_2 \prec_\alpha r_1$, $r_3 \prec_\alpha r_1$ et $r_5 \prec_\alpha r_4$. On peut voir à travers cet exemple qu'une règle dominante (selon la définition de la relation d'ordre $\prec_\alpha$) est plus informative qu'une règle dominée. En effet, selon r_2, la présence des motifs A et E dans la transaction (les objets) conduit toujours à la présence de l'item B. Autrement dit, si on veut que B soit dans la transaction, selon la règle r_2, la présence des motifs A et E sont nécessaires. Pourtant, selon r_1, c'est juste la présence du motif E qui est nécessaire à la présence de B, même logique pour r_4 et r_5. Pour trouver la semi-base positive exacte, il suffit de supprimer toutes les règles dominées. Le tableau 4.10 montre la différence entre $NBPE$ et le $SBPE$ avec un taux de réduction assez élevé (50%).

$NBPE$			Support
E	$\rightarrow$	B	2/3
AE	$\rightarrow$	B	1/2
DE	$\rightarrow$	B	1/3
C	$\rightarrow$	AB	1/2
CE	$\rightarrow$	AB	1/3
F	$\rightarrow$	ABE	1/3

$SBPE$			Support
E	$\rightarrow$	B	2/3
C	$\rightarrow$	AB	1/2
F	$\rightarrow$	ABE	1/3

Tableau 4.10 – $NBPE$ et $SBPE$ avec un $minSupp = 1/3$ et taux de réduction de 50%

Remarquons que la suppression des règles dominées entraine la perte des règles qui peuvent être dérivées à partir de ces règles dominées. Selon la proposition 4.14, les règles dérivées d'une règle dominée sont toujours moins informatives que la règle dominante. Donc, même si on supprime ces règles dominées, on ne risque en aucun cas de perdre d'information.

4.8.2 Semi-base positive approximative

L'idée est de considérer une base générée au niveau de confiance α (probabilité de ne pas se tromper) comme un ensemble d'informations (des règles) ayant les mêmes degré d'utilité ou d'importance. C'est à dire que la seule et unique condition pour qu'une règle soit considérée comme importante est d'avoir une mesure dépassant la valeur critique de M_{GK} correspondant ou tout simplement d'être dans la base (selon le niveau de confiance fixé par l'utilisateur). Dans ce cas, en utilisant la relation d'ordre « dominée par » définie dans l'ensemble des règles positives approximatives, il est possible de comparer certaines règles dans la base même si elles n'ont pas les mêmes mesures. Rappelons que dans la nouvelle base positive approximative $NBPA(\alpha)$ valide selon la mesure M_{GK}, une règle $r_2 : Z \rightarrow T$ est dominée par une autre règle $r_1 : X \rightarrow Y$, que l'on note par $r_2 \prec_\alpha r_1$ lorsque :

$$\begin{cases} M_{GK}(r_1) \geq M^{\alpha}_{GK}, \\ M_{GK}(r_2) \geq M^{\alpha}_{GK}, \\ Z \supseteq X \text{ et } T \subseteq Y\,. \end{cases}$$

La semi-base positive approximative ($SBPA(\alpha)$) n'est autre que la réduction de $NBPA(\alpha)$ en supprimant toutes les règles dominées par d'autres. Autrement dit, on garde seulement les règles dominantes. D'où la description ci-dessous :

$$SBPA(\alpha) = \{r \in NBPA(\alpha) \,/\, \nexists\, r' \in NBPA(\alpha)\,,\ r \prec_\alpha r'\}\,.$$

Exemple 14.
Prenons le contexte décrit dans le tableau $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$. *Avec un* $minSupp = 1/2$ *et une précision* $\alpha = 0,7$, *le tableau 4.11 nous montre les règles dans* $NBPA(\alpha)$ *et celles dans* $SBPA(\alpha)$, *avec un taux de réduction non négligeable (*29%*).*

	$NBPA$		
r_2	AB	$\rightarrow$	ABE
r_3	AB	$\rightarrow$	C
r_5	B	$\rightarrow$	AE
r_6	B	$\rightarrow$	AC
r_7	B	$\rightarrow$	E
r_8	B	$\rightarrow$	AB
r_9	B	$\rightarrow$	BD
r_{10}	A	$\rightarrow$	ABE
r_{11}	A	$\rightarrow$	BC
r_{12}	A	$\rightarrow$	AB
r_{13}	A	$\rightarrow$	AD
r_{14}	D	$\rightarrow$	BD
r_{15}	D	$\rightarrow$	AD
r_{20}	E	$\rightarrow$	ABE

	Sémi $-$ BPA		
r_2			
r_3			
r_5	B	$\rightarrow$	AE
r_6	B	$\rightarrow$	AC
r_7			
r_8	B	$\rightarrow$	AB
r_9	B	$\rightarrow$	BD
r_{10}	A	$\rightarrow$	ABE
r_{11}	A	$\rightarrow$	BC
r_{12}			
r_{13}	A	$\rightarrow$	AD
r_{14}	D	$\rightarrow$	BD
r_{15}	D	$\rightarrow$	AD
r_{20}	E	$\rightarrow$	ABE

Tableau 4.11 – Comparaison de $NBPA(0,7)$ et $SBPA(0,7) avec minSupp = 1/2$ et taux de réduction de 29%

4.8.3 Semi-base négative approximative

Rappelons que la nouvelle base négative approximative est constituée des règles entre les générateurs des motifs fermés. Bien qu'un générateur est toujours de taille réduite par rapport aux autres motifs non générateurs de sa classe, entre eux, comme les motifs fermés, les générateurs d'un contexte donné peuvent être comparables et par conséquent, les règles, elles aussi, peuvent être comparables selon la relation d'ordre « dominée par » définie dans l'ensemble des règles négatives approximatives. Prenons l'exemple des règles $r_1 : C \rightarrow \overline{BD}$ et $r_2 : C \rightarrow \overline{D}$ du contexte d'extraction $\mathcal{K}'$ décrit dans le tableau 4.9. Pour ces deux règles, nous avons : $\mathrm{M}_{GK}(C \rightarrow \overline{BD}) = 0,41$ et $\mathrm{M}_{GK}(C \rightarrow \overline{D}) = 0,56$. Au niveau de confiance $\alpha = 0,7$, les règles r_1 et r_2 sont valides. Selon la définition de la relation d'ordre dans l'ensemble des règles négatives approximatives, $r_1 : C \rightarrow \overline{BD}$ est dominée par la règle $r_2 : C \rightarrow \overline{D}$ (puisque $D \subset BD$). Donc, la règle $r_1 : C \rightarrow \overline{BD}$ et toutes ses dérivées dr, de la forme $X \rightarrow \overline{Y}$ avec $X \in [C]$ et $Y \in [BD]$ sont dominées par $r_2 : C \rightarrow \overline{D}$. En général, nous rencontrons une situation illustrée par la figure 4.10. Considérons trois fermés X, Y et Z de générateurs respectifs G_X, G_Y et G_Z avec G_Y, G_Z comparables (exp. $G_Y \subset G_Z$), en disposant de la relation d'ordre « dominée par » dans l'ensemble des règles négatives approximatives, nous sommes en mesure de comparer les deux règles r_1 et r_2. Dans le présent cas, la règle $r_1 : G_X \rightarrow \overline{G_Y}$ domine la règle $r_2 : G_X \rightarrow \overline{G_Z}$ et toutes ses dérivées.
Donc, au lieu de mettre les deux règles dans la base, on va se restreindre à r_1. C'est ainsi que nous avons défini la semi-base négative approximative :

$$SBNA(\alpha) = \{r \in NBNA(\alpha) \,/\, \nexists\, r' \in NBNA(\alpha),\ r \prec_\alpha r'\}.$$

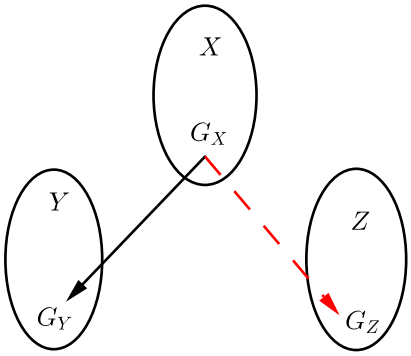


FIGURE 4.10 – Choix des règles négatives dominantes

Comme dans le cas des autres semi-bases, certaines règles « redondantes » et leurs règles dérivées seront supprimées de la nouvelle base des règles négatives approximatives pour former la semi-base négative approximative. La disponibilité de ce type de base permettra à l'utilisateur de focaliser ses interprétations dans l'ensemble des règles les plus informatives seulement. En cas de besoin, si l'utilisateur souhaite avoir les règles jugées redondantes, il est toujours possible de faire appel à $NBNA$.

4.8.4 Semi-base négative exacte

Proposition 4.17. *Soient $r_1 : X_1 \rightarrow \overline{Y_1}$ et $r_2 : X_2 \ \overline{Y_2}$ deux éléments de la Nouvelle Base des règles Négatives exactes d'un certain contexte $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$.*
Par rapport à la relation d'ordre d'inclusion, nous avons les deux résultats suivants :

1. *X_1 et X_2 ne sont pas comparables.*
2. *Y_1 et Y_2 ne sont pas comparables.*

Preuve. Prenons deux éléments quelconque $r_1 : X_1 \rightarrow \overline{Y_1}$ et $r_2 : X_2 \ \overline{Y_2}$ de la Nouvelle Base des règles Négatives Exactes ($NBNE$). Comme r_1 et r_2 sont des éléments de $NBNE$, il existe deux motifs fermés de supports non nuls F_1 et F_2 tels que X_1 soit un générateur de F_1 et X_2 soit un générateur de F_2, de plus, F_1 et F_2 sont des éléments de $Bd^+(0)$. Désignons par γ l'opérateur de fermeture associé au contexte $\mathcal{K}$. Effectuons un raisonnement par l'absurde. Supposons que $X_1 \subset X_2$, comme l'opérateur de fermeture γ est isotone, $\gamma(X_1) \subset \gamma(X_2)$, donc $F_1 \subset F_2$. Autrement dit, F_2 est un sur-ensemble de F_1 et ceci est en contradiction avec le fait que F_1 est élément de $Bd^+(0)$, donc en contradiction avec le fait que $r_1 : X_1 \rightarrow Y_1$ est élément de $NBNE$. On peut donc conclure que X_1 ne peut pas être un sous-ensemble de X_2. Avec le même raisonnement, nous pouvons montrer que X_2 ne peut pas être un sous ensemble de X_1. X_1 et X_2 ne sont pas comparables.
D'autres part, les conséquents des règles dans $NBNE$ sont des singletons. Y_1 et Y_2 ne peuvent donc pas être comparables. □

Corollaire 4.5. *Aucune relation d'ordre basée sur la relation d'inclusion ensembliste ne peut être définie dans $NBNE$. C'est à dire qu'en se servant uniquement de la relation d'inclusion, il est impossible d'ordonner les éléments de $NBNE$ (les règles ne peuvent pas être comparables). Donc, aucune semi-base ne sera définie à partir de $NBNE$.*

4.9 Conclusion partielle

Après avoir apporté des critiques par rapport à la construction des anciennes bases des règles M_{GK}-valides, nous avons pu proposer des nouvelles bases générées à partir de l'utilisation des motifs fermés, des générateurs et les bordures positives. Dans la construction de ces nouvelles bases, nous avons aussi privilégié l'utilisation de quelques propriétés souhaitables des règles non redondantes. Parmi ces propriétés, nous pouvons citer : la minimalité des prémisses, la maximalité des conséquents des règles positives et la disjonction entre une prémisse et un conséquent d'une règle. En ce qui concerne les règles négatives, nous avons montré qu'entre deux ou plusieurs règles négatives valides, celles qui ont des motifs négatifs minimaux sont les plus informatives (dans la prémisse comme dans le conséquent).
En tenant compte de ces propriétés qui influencent l'informativité des règles d'association, nous avons pu proposer quatre nouvelles bases des règles valides selon la mesure M_{GK} : Nouvelle Base des règles Positives Exactes ($NBPE$), Nouvelle Base des règles Négatives Exactes ($NBNE$), Nouvelle Base des règles Positives Approximatives ($NBPA$), Nouvelle Base des règles Négatives Approximatives ($NBNA$). Nous avons proposé des relations d'ordre dans chaque catégorie des bases des règles M_{GK}-valides. Ces relations d'ordre nous permettront de comparer les règles entre elles. Ainsi, nous pouvons déterminer si une règle est plus informative que d'autres. En exploitant cette notion d'ordre dans l'ensemble des règles, nous avons pu réduire certains éléments des nouvelles bases pour construire des ensembles des règles que nous avons appelé semi-bases.
Relativement à un niveau de confiance fixé par l'utilisateur, les semi-bases regroupent toutes les règles les plus informatives seulement. De plus, il est appréciable que la considération des semi-bases peut engendrer une réduction notable du nombre des règles en passant d'une base vers la semi-base associée (cf. Tableau 4.10 et Tableau 4.11). Le mot semi-base vient du fait que ces ensembles ne sont pas tout à fait des bases parce qu'avec eux, il est impossible de retrouver certaines règles valides. Nous avons donc pu construire trois semi-bases des règles M_{GK}-valides : Semi-Base Positive Exacte ($SBPE$), Semi-Base Positive Approximative ($SBPA$) et Semi-Base Négative Approximative ($SBNA$).
Dans le but d'automatiser l'extraction des connaissances à partir des données en utilisant les outils mathématiques que nous venons de décrire, nous allons consacrer le prochain chapitre à la conception des algorithmes relatifs à la génération des bases des règles M_{GK}-valides.

Chapitre 5

Algorithmes d'extraction des bases des règles M_{GK} valides

Sommaire

5.1 Introduction

Il est évidemment impensable de traiter des données réelles avec plusieurs dizaines, voire une centaine des variables sans l'aide des outils informatiques. De nombreux logiciels sont actuellement disponibles pour extraire des connaissances à partir des données (ECD), notamment, l'extraction des règles d'association. On peut citer, entre autres, le package arules du logiciel R, dans lequel on peut se servir de plusieurs mesures comme le support, la confiance, le lift, etc. Il y a aussi le logiciel CHIC et R-CHIC, développé par l'équipe de Régis GRAS pour l'extraction des règles d'association basées sur l'intensité d'implication, la mesure que nous avons utilisée dans l'analyse des données issues de l'expérimentation décrite au chapitre 2. Étant donné les limites du couple support et confiance détaillées au chapitre 3, et le fait que CHIC et R-CHIC n'étaient pas conçus pour mesurer des règles négatives, et surtout, les propriétés mathématiques et sémantiques (compatibilité à la logique classique, voir § 2.9) de la mesure M_{GK}, il est légitime de penser à la conception d'un outil d'extraction des connaissances à partir des données basé sur la mesure M_{GK}. Par rapport à cet objectif, nous

allons consacrer ce chapitre à la description des algorithmes permettant de faire une ECD basée sur la mesure M_{GK}. Précisons avant tout que les nouvelles bases M_{GK}-valides sont composées essentiellement des règles entre un fermé et/ou un générateurs. Par conséquent, pour nos algorithmes, il est nécessaire de disposer de ces types de motifs. Plusieurs travaux sur l'extraction des fermés et des générateurs ont été effectués. On peut citer entre autres les travaux de (Bastide *et al.*, 2000 ; Le Floc'h *et al.*, 2003 ; Salleb, 2003 ; Hamrouni *et al.*, 2005 ; Nguifo, 2004 ; Hafida, 2013). Nous allons considérer ces études pour l'extraction des fermés et générateurs. Autrement dit, pour un contexte d'extraction $\mathcal{K}$ donné, nous supposons que les fermés et les générateurs sont disponibles et, à partir de ces motifs, nous allons proposer des algorithmes d'extraction des bases des règles M_{GK}-valides.

5.2 Bases des règles positives

Dans cette section, nous allons nous concentrer sur l'extraction des bases des règles positives. Considérons donc un contexte binaire d'extraction $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ et, compte tenu des hypothèses faites ci-dessus, désignons par FF_k l'ensemble des fermés fréquents (relativement à un *minsup* fixé) de taille k ($1 \leq k \leq n$), avec n désignant la taille des données ou le nombre des transactions ou encore, le nombre des objets du contexte ($n = \text{card}(\mathcal{O})$). Ensuite, désignons par $\mathcal{G}_F$ l'ensemble des générateurs d'un fermé F. Rappelons qu'une règle est dite positive lorsque les composantes (prémisse et conséquent) sont des motifs positifs, c'est à dire, la prémisse et conséquent sont des éléments de $\mathscr{P}(\mathcal{I})$. Il faut juste s'assurer, dans la mesure du possible, que la prémisse et le conséquent soient disjoints.

5.2.1 Bases des règles positives exactes

Selon la définition 4.5, la nouvelle base des règles positives exactes est formée par des règles ayant des prémisses et conséquents d'une même classe. Il suffit donc de connaître l'ensemble des fermés F et pour chacun d'eux, on prend leurs générateurs G ($G \neq F$) comme prémisse et la différence $F \setminus G$ comme conséquent. Étudions le cas particulier du motif $\emptyset$ lorsqu'il n'est pas fermé. Si $\emptyset$ n'est pas fermé, alors il existe un motif non vide X tel que $\gamma(\emptyset) = X$ ($\emptyset$ est alors un pseudo-fermé). Cette situation ne peut être réalisée que lorsque le motif X, et évidemment, tous ses sous-motifs sont présents dans toutes les transactions. Les règles $\emptyset \rightarrow X$ ($\emptyset \rightarrow \gamma(\emptyset)$) et toutes les règles $\emptyset \rightarrow Y$, avec $Y \subseteq X$, ont des Confiances (probabilité conditionnelle) égales à 1 et M_{GK} égales à 0. En effet,

$$\begin{aligned} Conf\left(\emptyset \rightarrow X\right) &= P\left(X'/\emptyset'\right) \\ &= P\left(X'/\mathcal{O}\right) \\ &= \frac{P(X' \cap \mathcal{O})}{P(\mathcal{O})} \\ &= 1. \end{aligned}$$

Comme $P(\emptyset') = P(\mathcal{O}) = 1$, les motifs $\emptyset$ et X sont indépendants.

$$\begin{aligned} M_{GK}\left(\emptyset \rightarrow X\right) &= \frac{P\left(X'/\emptyset'\right) - P\left(X'\right)}{P\left(X'\right)} \\ &= \frac{1-1}{1} \\ &= 0. \end{aligned}$$

Comme l'a fait remarquer (Pasquier, 2000a), ces règles n'apportent aucune information supplémentaire par rapport au motif $\gamma(\emptyset)$ dont le support est égal à 1. Remarquons que le motif X est fermé ($\gamma(\gamma(\emptyset)) = \gamma(\emptyset)$) et toujours fréquent (Supp(X)=Card($\mathcal{O}$)). Pour éviter ce type de règle, nous devons enlever le motif X de l'ensemble des fermés fréquents. Nous allons maintenant décrire un algorithme d'extraction de base des règles positives exactes.

Algorithme 3 Base des Règles Positives Exactes

Entrée : $FF_{\mathcal{K}}$ Ensemble des fermés fréquents
$\mathcal{G}_F$ Ensemble des générateurs d'un fermé F
Sortie : $NBPE$ Nouvelle Base Positive Exacte

1: $X = \emptyset$
2: **Pour** Chaque i dans $\mathcal{I}$ **faire**
3: **Si** (Supp(i) == n) **alors**
4: $X = X \cup \{i\}$
5: **Fin Si**
6: **Fin Pour**
7: $FF_{\mathcal{K}} = FF_{\mathcal{K}} \setminus X$
8: $NBPE = \emptyset$
9: **Pour** Chaque F de $FF_{\mathcal{K}}$ **faire**
10: **Pour** Chaque G de $\mathcal{G}_F$ **faire**
11: **Si** ($\gamma(G) \neq G$) **alors**
12: $NBPE = NBPE \cup (\{G \rightarrow F \setminus G\}, \text{Supp}(G))$
13: **Fin Si**
14: **Fin Pour**
15: **Fin Pour**
16: Retourner $NBPE$

L'algorithme d'extraction de la Nouvelle Base des règles Positives Exactes ($NBPE$) commence par la détection du motif fermé X dont le support est égal au nombre de transactions, c'est à dire que X est présent dans toutes les transactions (ligne 1 à 5). Dans cette situation, $\gamma(\emptyset) = X$, le motif $\emptyset$ est donc l'unique générateur de X. En dehors du fait de savoir que X est présent dans toutes les transactions, l'implication $\emptyset \rightarrow X$ n'apporte pas d'information utile. De plus, le motif X ne sera jamais le conséquent d'une quelconque règle positive exacte valide. Il n'y a donc aucun intérêt de garder X dans l'ensemble des motifs fermés fréquents. Il convient donc de le supprimer avant l'extraction de la $NBPE$ (ligne 7). Ensuite, on examine un par un les motifs fermés, et pour chacun des fermés F, on prend leurs générateurs G qui sont différents de F et formés des règles de type $G \rightarrow F \setminus G$ (ligne 10 à 15).

Exemple 15.
Prenons l'exemple du contexte décrit dans le tableau 2.1. L'algorithme doit parcourir l'ensemble des items pour vérifier s'il en existe ceux de support égales à 6 (Card($\mathcal{O}$)).

Dans cet exemple, aucun item n'est présent dans toutes les transactions (les objets). Nous allons maintenant parcourir l'ensemble $FF_{\mathcal{K}}$ (fermés Fréquents avec un minSupp = 1/3). Pour chacun de ces fermés, l'algorithme examine leurs générateurs et on construit les règles exactes à partir des motifs fermés F et son générateur G_F lorsque $G_F \neq F$. Le tableau 5.2

Items	A	B	C	D	E	F
Supports	5	5	3	4	4	2

Tableau 5.1 – Test de Support des Items

montre les étapes de construction de la nouvelle base positive exacte en exécutant l'algorithme 3.

Fermés	Générateurs	$NBPE$	Fermés	Générateurs	$NBPE$
A	A		BE	E	$E \rightarrow B$
B	B		ABE	AE	$AE \rightarrow B$
D	D		BDE	DE	$DE \rightarrow B$
AB	AB		ABC	C	$C \rightarrow AB$
BD	BD		$ABEF$	F	$F \rightarrow ABE$
AD	AD		$ABCE$	CE	$CE \rightarrow AB$
ABD	ABD				

Tableau 5.2 – Génération de $NBPE$

5.2.2 Semi-base des règles positives exactes

Nous avons montré au paragraphe 4.8.1 que les fermés et les générateurs peuvent être comparables. Pour avoir des règles plus informatives, nous avons défini la relation d'ordre « *dominé par* » dans l'ensemble des règles exactes. Dans un contexte quelconque $\mathcal{K}$, on a :

$$r_1 \prec r_2 \Leftrightarrow \begin{cases} M_{GK}(r_1) = M_{GK}(r_2), \\ X_1 \supseteq X_2 \; et \; Y_1 \subseteq Y_2. \end{cases}$$

À partir de cette relation d'ordre, nous avons défini l'ensemble $SBPE$:

$$SBPE = \{r : G_F \rightarrow F \backslash G_F, \; telle \; que \; F \in FF_{\mathcal{K}}, \; G_F \in \mathcal{G}_F \; et \; \nexists r' / \; r \prec r'\}.$$

Nous allons faire une description d'un algorithme d'extraction de $SBPE$. L'idée est de partir de la $NBPE$ et supprimer les règles qui sont dominées par d'autres.
La définition des semi-bases des règles est fondée sur la comparaison, selon la relation d'ordre « $\prec$: dominée par », des règles valides. Nous allons donc définir une fonction nommée CompE (comparaison des règles Exactes) qui prend deux règles valides comme arguments et retourne la règle dominante si les deux règles sont comparables, l'ensemble vide dans le cas contraire. Son expression est :

$$CompE(r_1, \; r_2) = \begin{cases} r_1 & si \quad r_2 \prec r_1, \\ r_2 & si \quad r_1 \prec r_2, \\ \emptyset & sinon. \end{cases}$$

La valeur de la fonction CompE dépend certainement de la définition de la relation d'ordre « dominée par ». Nous allons décrire un algorithme qui peut fournir la valeur de CompE.

Algorithme 4 Fonction CompE

Entrée : r_1, r_2 deux règles
Sortie : r règle dominante ou $\emptyset$

Si ($r_1.Premisse \subseteq r_2.Premisse$ ET $r_2.Consequent \subseteq r_1.Consequent$) **alors**
  $r = r_1$
Sinon
  Si ($r_2.Premisse \subseteq r_1.Premisse$ ET $r_1.Consequent \subseteq r_2.Consquent$) **alors**
    $r = r_2$
  Sinon
    $r = \emptyset$
  Fin Si
Fin Si
Retourner r

À partir de la définition et de l'algorithme de la fonction CompE, on peut facilement décrire un algorithme d'extraction de Semi-Base des règles Positives Exactes.

Algorithme 5 Semi-Base des Règles Positives Exactes

Entrée : $NBPE$
Sortie : $SBPE$

1: $SBPE = NBPE$
2: **Pour** Chaque r_1 de $NBPE$ **faire**
3: EspaceTest = $SBPE \setminus r_1$
4: **Pour** Chaque r_2 dans EspaceTest **faire**
5: r =CompE(r_1, r_2)
6: **Si** ($r == r_1$) **alors**
7: $SBPE = SBPE \setminus r_2$
8: **Fin Si**
9: **Si** ($r == r_2$) **alors**
10: $SBPE = SBPE \setminus r_1$
11: EspaceTest = $\emptyset$
12: **Fin Si**
13: **Fin Pour**
14: **Fin Pour**
15: Retourner $SBPE$

Puisque $SBPE$ est un sous-ensemble de la $NBPE$, l'algorithme 5 commence par l'initialisation de $SBPE$ par les éléments de $NBPE$. Ensuite, il faut supprimer les règles qui sont dominées par d'autres règles. Pour chacune des règles dans $NBPE$ (ligne 2), on la compare avec les restes des règles exactes qui sont dans $SBPE$. S'il existe une règle r_2 qui domine la règle r_1 (ligne 9 à 12), alors r_1 sera supprimée de l'ensemble $SBPE$, ensuite, on recommence avec une nouvelle règle de $NBPE$ (ligne 11). Par contre, s'il existe une règle r_2 de l'espace test (EspaceTest) qui est dominée par r_1, alors r_2 doit être enlevée de la $SBPE$ et on continue d'explorer l'espace test (parce qu'il est possible de trouver d'autres r_2 qui soient comparables à r_1). L'algorithme s'arrête lorsque toutes les règles dans $NBPE$ sont testées.

5.2.3 Bases des règles positives approximatives

Nous allons décrire un algorithme d'extraction de nouvelle base des règles positives approximatives. Selon la définition 4.9, la description de $NBPA$ nécessite la disponibilité de l'ensemble noté R_{XY}, représentantes des règles de prémisse dans $[X]$ et de conséquent dans $[Y]$ et ceci, pour chaque couple de fermés fréquents (X, Y). Appelons donc par $RR(X, Y)$, une fonction permettant de nous donner l'ensemble R_{XY} (pour tous X, Y dans l'ensemble des fermés fréquents). La fonction $RR(X, Y)$ est décrit par l'algorithme 2. L'algorithme a besoin d'une liste $Khi2(\alpha)$, contenant les valeurs théoriques de χ^2 à un degré de liberté et avec les seuils les plus utilisés. Le tableau 5.3 résume les significations de certaines variables utilisées dans l'algorithme 6.

n, n_X, n_Y, n_{XY}	Désignent respectivement le support de $\mathcal{O}, X, Y, X \cup Y$
MCC	Ensemble des Motifs Candidats Conséquents
$Khi2(\alpha)$	Valeur théorique de χ^2 à 1 ddl et au niveau de confiance α (risque d'erreur $1 - \alpha$)

Tableau 5.3 – Quelques variables utilisées dans l'algorithme 6

Algorithme 6 Base des Règles Positives Approximatives

Entrée : $FF_{\mathcal{K}}$ Ensemble des Fermés Fréquents
α : Niveau de confiance utilisé pour le calcul de valeur théorique de χ^2
Sortie : $NBPA$ Base positive Approximative

1: $Khi2 = Khi2(\alpha)$
2: **Pour** Chaque X dans $FF_{\mathcal{K}}$ **faire**
3: **Pour** Chaque Y dans $FF_{\mathcal{K}}$ **faire**
4: **Si** $(Y \subseteq X)$ **alors**
5: $MCC = FF_{\mathcal{K}} \setminus Y$
6: **Fin Si**
7: **Fin Pour**
8: **Pour** Chaque Y dans MCC **faire**
9: $n_X = Card(X')$
10: $n_Y = Card(Y')$
11: $n_{XY} = Card(X' \cap Y')$
12: $Ecart = n_{XY}/n_X - n_Y/n$
13: **Si** $Ecart > 0$ **alors**
14: mgk $= Ecart/(1 - n_Y/n)$
15: mgkcr $= \sqrt{\frac{1}{n}\frac{n-n_X}{n_X}\frac{n_Y}{n-n_Y}Khi2}$
16: **Si** mgk $\geq$ mgkcr **alors**
17: $R_{XY} = RR(X, Y)$
18: $NBPA = NBPA \cup \{R_{XY}\}$
19: **Fin Si**
20: **Fin Si**
21: **Fin Pour**
22: **Fin Pour**
23: Retourner BPA

Connaissant l'ensemble des fermés fréquents $FF_{\mathcal{K}}$ avec leurs supports et le niveau de confiance α, choisi par l'utilisateur pour le calcul de valeur critique de χ^2, l'algorithme 6 teste toutes les règles $X \rightarrow Y$, avec $X, Y \in FF_{\mathcal{K}}$. Pour une prémisse X de $FF_{\mathcal{K}}$, l'algorithme 6 commence par déterminer les motifs qui sont susceptibles d'être conséquents du motif X. Nous les avons appelé Motifs Candidats Conséquents (MCC) (ligne 3 à 7). Pour chacune des Y dans MCC, on détermine si le motif Y favorise le motif X en testant le signe de $P(Y'/X') - P(Y')$ (ligne 9 à 13). Dans le cas où cet écart $(P(Y'/X') - P(Y'))$ est positif, l'algorithme évalue la valeur de M_{GK} et celle de M^{α}_{GK} de la règle $X \rightarrow Y$. Ces deux quantités permettent ensuite de déterminer si la règle $X \rightarrow Y$ est valide ou non. Dans le premier cas où $X \rightarrow Y$ est valide, on fait appel à l'algorithme 2 pour déterminer l'ensemble R_{XY}, contenant la ou les représentantes des règles de prémisse dans la classe de X ($[X]$) et de conséquent dans la classe de Y ($[Y]$) (ligne 8 à 17). Ce sont les éléments de R_{XY} qui vont constituer l'ensemble des règles positives approximatives valides selon la mesure M_{GK} (ligne 18). Nous allons dérouler l'algorithme 6 avec les données du contexte $\mathcal{K}$ décrit dans le tableau 2.1.

$X = A$

Y	EC	M_{GK}	M^{α}_{GK}
B	-0.03	×	×
D	-0.06	×	×
AD	0.1	0.20	0.21
BD	-0.10	×	×
AB	0.13	0.40	0.30
BE	-0.06	×	×
ABC	0.10	0.20	0.21
ABE	0.10	0.20	0.21
$R_{XY} = \{A \rightarrow AB\}$			

$X = B$

Y	EC	M_{GK}	M^{α}_{GK}
A	-0.03	×	×
D	-0.06	×	×
AD	-0.10	×	×
BD	0.10	0.20	0.21
AB	0.13	0.40	0.30
BE	0.13	0.40	0.30
ABC	0.10	0.20	0.21
ABE	0.10	0.20	0.21
$R_{XY} = \{B \rightarrow AB, B \rightarrow E\}$			

$X = D$

Y	EC	M_{GK}	M^{α}_{GK}
B	-0.08	×	×
A	-0.08	×	×
AD	0.25	0.50	0.33
BD	0.25	0.50	0.33
AB	-0.16	×	×
BE	-0.16	×	×
ABC	-0.25	×	×
ABE	-0.25	×	×
$R_{XY} = \{D \rightarrow AD, D \rightarrow BD\}$			

$X = AD$

Y	EC	M_{GK}	M^{α}_{GK}
B	-0.16	×	×
BD	0.16	0.33	0.47
AB	0	×	×
BE	-0.33	×	×
ABC	-0.16	×	×
ABE	-0.16	×	×
$R_{XY} = \{\}$			

$X = BD$

Y	EC	M_{GK}	M^{α}_{GK}
A	-0.16	×	×
AD	0.16	0.33	0.47
AB	0	×	×
BE	0	×	×
ABC	-0.16	×	×
ABE	-0.16	×	×
$R_{XY} = \{\}$			

$X = AB$

Y	EC	M_{GK}	M^{Cr}_{GK}
D	-0.16	×	×
AD	0	×	×
BD	0	×	×
BE	0.08	0.25	0.47
ABC	0.25	0.50	0.33
ABE	0.25	0.50	0.33
$R_{XY} = \{AB \rightarrow C, AB \rightarrow ABE\}$			

$X = BE$

Y	EC	M_{GK}	M^{α}_{GK}
D	-0.16	×	×
A	-0.08	×	×
AD	-0.25	×	×
BD	0	×	×
AB	0.08	0.25	0.47
ABC	0	×	×
ABE	0.25	0.50	0.33
$R_{XY} = \{BE \rightarrow ABE\}$			

$X = ABC$

Y	EC	M_{GK}	M^{α}_{GK}
D	-0.33	×	×
AD	-0.16	×	×
BD	-0.16	×	×
BE	0	×	×
ABE	0.16	0.33	0.47
$R_{XY} = \{\}$			

$X = ABE$

Y	EC	M_{GK}	M^{α}_{GK}
D	-0.33	×	×
AD	-0.16	×	×
BD	-0.16	×	×
ABC	0.16	0.33	0.47
$R_{XY} = \{\}$			

Tableau 5.4 – Étapes d'exécution de l'algorithme 6

En prenant un $minSupp = 1/2$ et un niveau de confiance $\alpha = 0,7$, les 9 tableaux groupés sous le nom tableau 5.4 montrent un exemple des étapes de construction de BPA par l'algorithme 6. Les motifs conséquents candidats Y ($Y \in MCC$) sont rangés en lignes et pour

une prémisse X fixée (X un motif fermé fréquent de l'ensemble noté $FF_{\mathcal{K}}$), on teste l'écart à l'indépendance $P(Y'/X') - P(Y')$ (colonne EC). Pour un couple de motif X, Y, ($X \in FF_{\mathcal{K}}$ et $Y \in MCC$), si l'écart est positif, alors on évalue la valeur de M_{GK} et celle de M^{α}_{GK}. Une fois qu'une règle $X \to Y$ est valide ($\mathrm{M}_{GK}(X \to Y) \geq \mathrm{M}^{\alpha}_{GK}$), on fait appel à l'algorithme 2 pour trouver les représentantes des règles de prémisse dans $[X]$ et de conséquent dans $[Y]$. D'où la base positive approximative du contexte décrit dans le tableau 2.1 au niveau de confiance $\alpha = 0,75$:

$$\begin{aligned} BPA \;=\; & \{A \to AB, B \to AB, B \to E, D \to AD, D \to BD, \\ & AB \to C, AB \to ABE, BE \to ABE\}. \end{aligned}$$

5.2.4 Semi-base des règles positives approximatives

À partir de la relation d'ordre « dominé par », définie dans l'ensemble des règles positives approximatives, nous avons pu définir la semi-base positive approximative. La semi-base positive approximative est constituée des règles dominantes. Rappelons que cette notion de domination dans l'ensemble des règles positives a un sens lorsque l'utilisateur accorde les mêmes intérêts et, au même pied, toutes les règles valides relativement au niveau de confiance α qu'il a fixé. Pour mieux comprendre l'intérêt de la notion de semi-base des règles positives approximatives, prenons un exemple très formel. Considérons les trois règles $r_1 : A \to BCD, r_2 : AB \to CD$ et $r_3 : A \to BCDE$. Supposons que $\mathrm{M}_{GK}(r_1) = 0.71, \mathrm{M}_{GK}(r_2) = 0.70$ et $\mathrm{M}_{GK}(r_3) = 0.69$ et qu'au niveau de confiance α, les trois règles sont valides ($\mathrm{M}_{GK}(r_1) \geq \mathrm{M}^{\alpha}_{GK}(r_1), \mathrm{M}_{GK}(r_2) \geq \mathrm{M}^{\alpha}_{GK}(r_2)$ et $\mathrm{M}_{GK}(r_3) \geq \mathrm{M}^{\alpha}_{GK}(r_3)$). Ici, les mesures M_{GK} de chacune des règles r_1, r_2 et r_3 sont certes différentes, mais d'un autre côté, les mesures sont « voisines » les unes des autres (voisines dans le sens où leurs différences sont petites). De plus, toutes les trois dépassent leurs valeurs critiques respectives. Donc les trois règles sont valides au niveau de confiance α. Passons maintenant à l'interprétation. On peut remarquer que la mesure de r_1 est plus élevée que les mesures des deux autres règles mais c'est r_3 qui apporte beaucoup plus d'information que r_1 et r_2, puisque, par rapport à ces deux dernières règles, r_3 a une prémisse minimale et conséquent maximal. Autrement dit, les informations véhiculées par r_1 et r_2 sont déjà incluses dans les informations apportées par r_3. Compte tenu des éventuelles quantités trop importantes des règles valides dont la plupart apportent des informations moins générales que d'autres, on peut très bien imaginer des situation où l'utilisateur s'intéresse aux règles plus informatives au niveau de confiance fixé. C'est dans ces situations que la notion de semi-base positive approximative prend son sens et son importance. En ce qui concerne l'algorithme d'extraction de semi-base positive approximative, il ne diffère qu'à peu de chose près de l'algorithme d'extraction de semi-base positive exacte. En effet, la différence se trouve au niveau de la définition de la relation d'ordre « dominé par » donc, au niveau de l'algorithme permettant de comparer deux règles positives approximatives valides. Après avoir rappelé la définition de la relation d'ordre « dominé par » dans l'ensemble des règles positives approximatives, nous allons décrire un algorithme permettant de comparer deux règles positives approximatives valide au niveau de confiance α.
Selon la définition donnée au paragraphe 4.8.2, une règle $r_1 : X_1 \to Y_1$ est dominée par une

autre règle r_2 au niveau de confiance α lorsque :

$$\begin{cases} M_{GK}\,(r_1) \geq M^{\alpha}_{GK}, \\ M_{GK}\,(r_2) \geq M^{\alpha}_{GK}, \\ X_1 \supseteq X_2 \; et \; Y_2 \subseteq Y_1\,. \end{cases}$$

À titre d'exemple, prenons le cas des deux règles $r_5 : B \rightarrow AE$ et $r_7 : B \rightarrow E$. D'abord, au niveau de confiance $\alpha = 0,70$ et en prenant un $minSup = 1/2$, les deux règles sont valides (cf. tableau 4.11). Dans le présent cas, nous avons :

$$r_5.premisse = r_7.premisse \text{ et } r_7.consequent \subseteq r_5.consequent.$$

Bien que $\mathrm{M}_{GK}(r_5) \neq \mathrm{M}_{GK}(r_7)$, les deux règles sont valides au niveau de confiance $\alpha = 0,70$. Pour un utilisateur qui considère le niveau de confiance α comme l'unique balise de validité de la décision qu'il va prendre (il accorde le même ordre d'importance à toutes les règles valides à ce niveau), la règle r_7 n'apporte aucune information supplémentaire à l'égard de la règle r_5. On doit montrer que la règle r_7 est dominée par la règle r_5, et on note $r_7 \prec r_5$. Dans cette situation, cet utilisateur peut se restreindre à r_5, du moins en première interprétation. Après, si le nombre des règles valides n'est pas « trop élevé » ou que l'on souhaite avoir des idées sur l'« intensité d'implication » de toutes les règles valides, on peut toujours faire appel à $NBPA$.
Étant donné qu'à un niveau de confiance α fixé, toutes les règles dans $NBPA$ sont valides (de mesure dépassant leur valeur critique), en partant de ces éléments, nous pouvons avoir le même algorithme de comparaison des règles que celui des règles positives exactes (algorithme 4). De plus, comme dans l'ensemble des règles positives exactes, dans l'ensemble des règles positives approximatives, la construction de semi-base est aussi fondée sur le principe de minimalité de prémisse et maximalité des conséquents. Par rapport à cela, le même algorithme d'extraction de semi-base positive exacte (algorithme 5), appliqué à $NBPA$, fournira l'ensemble que nous avons appelé semi-base positive approximative.

5.3 Génération des règles négatives à partir des motifs positifs

Nous avons vu au chapitre 3 qu'on a généré les bases des règles (positives et négatives) à partir des fermés fréquents. Nous avons vu aussi, au paragraphe 3.3.2, qu'avec l'algorithme 1, la génération des règles négatives à partir des motifs positifs peut conduire à l'obtention de règles négatives non valides (constituées des motifs non fréquents) ou à la perte de règles négatives valides (constituées des motifs négatifs dont les motifs positifs associés sont non fréquents). Pour éviter ce type de problème, avant de proposer les nouveaux algorithmes d'extraction des bases des règles M_{GK}-valides, nous allons analyser de près la position par rapport à un $minSupp$ donné, du support d'un motif $\overline{Y}$ connaissant celui du motif Y. Cette analyse nous permettra de déduire la validité d'une règle de type $X \rightarrow \overline{Y}$ à partir de la description des motifs positifs X et Y. Dans la pratique, on s'intéresse souvent aux règles constituées des motifs « représentatifs ». Le concept sur le support de motif est justement utilisé pour quantifier cette représentativité. Autrement dit, pour un $minSupp$ fixé, on s'intéresse aux règles constituées des motifs ayant le support plus grand que le $minSupp$. On les

appelle motifs fréquents. Comme il est très pratique de générer les règles négatives à partir des motifs positifs, nous allons montrer, à travers l'énumération de tous les cas possibles, la validité des règles de type $X \to \overline{Y}$ selon la fréquence des motifs X et Y.
Rappelons que, selon la propriété 3.6 et la proposition 4.1, si X défavorise Y (X favorise $\overline{Y}$), alors :

$$\begin{cases} M_{GK}\left(X \to \overline{Y}\right) &= M_{GK}\left(Y \to \overline{X}\right), \\ M^{\alpha}_{GK}\left(X \to \overline{Y}\right) &= M^{\alpha}_{GK}\left(Y \to \overline{X}\right). \end{cases}$$

Donc, si on ne se réfère qu'à la mesure M_{GK} et sa valeur critique, on peut déduire la validité de $Y \to \overline{X}$ à partir de celle de $X \to \overline{Y}$. Or, selon la définition 4.2, le test de validité d'une règle est assuré par les deux mesures Support et M_{GK}. Donc, si $X \to \overline{Y}$ est valide au niveau de confiance α, alors on a :

$$\begin{cases} M_{GK}\left(X \to \overline{Y}\right) \geq M^{\alpha}_{GK}\left(X \to \overline{Y}\right), \\ \mathrm{Supp}\left(X\right) \geq minSupp, \\ \mathrm{Supp}\left(\overline{Y}\right) \geq minSupp. \end{cases}$$

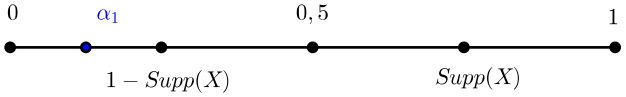
À partir de la validité de $X \to \overline{Y}$, peut-on systématiquement déduire celle de $Y \to \overline{X}$? La réponse est négative. En effet, on a :

$$\begin{cases} \mathrm{Supp}\left(\overline{X}\right) = 1 - \mathrm{Supp}\left(X\right), \\ \mathrm{Supp}\left(\overline{Y}\right) = 1 - \mathrm{Supp}\left(Y\right). \end{cases}$$

Le fait que X et $\overline{Y}$ sont des motifs fréquents ne constitue pas une raison suffisante pour prononcer sur la fréquence de $\overline{X}$ et Y.
Donc, on peut constater que si $M_{GK}\left(X \to \overline{Y}\right) \geq M^{\alpha}_{GK}\left(X \to \overline{Y}\right)$, $\mathrm{Supp}\left(\overline{X}\right) \geq minSupp$ et $\mathrm{Supp}\left(\overline{Y}\right) \geq minSupp$, alors on peut seulement affirmer que $M_{GK}\left(Y \to \overline{X}\right) \geq \mathrm{M}^{\alpha}_{GK}$ mais on ne peut rien dire sur la position de $\mathrm{Supp}(Y)$ et $\mathrm{Supp}(\overline{X})$ par rapport au $minSupp$, sinon que $Supp(\overline{X}) \leq 1 - minSupp$. Appuyons ces propos par des exemples énumérant de manière exhaustive tous les cas possibles. Nous allons distinguer deux cas selon le signe de la différence entre $\mathrm{Supp}(X)$ et $1 - \mathrm{Supp}(X)$.

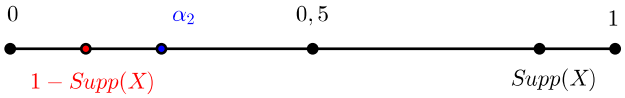
***Premier cas : Supp*$(X) > 0,5$**
Lorsque $\mathrm{Supp}(X) > 0,5$, la différence $\mathrm{Supp}(X) - (1 - \mathrm{Supp}(X))$ est positive. Donc, $\mathrm{Supp}(X)$ est plus grand que $1 - \mathrm{Supp}(X)$. Trois éventualités sont envisageables. Nous allons montrer à partir des exemples que pour un motif fréquent X, le motif négatif $\overline{X}$ ne l'est pas toujours.

— **Exemple 1 :**
Prenons un $minSupp = 0,3$ et un motif X tel que $\mathrm{Supp}(X) = 0,6$.



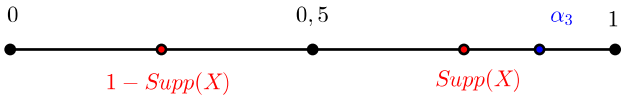
Pour cet item X, le motif négatif $\overline{X}$ est aussi un motif fréquent. En effet, si nous calculons son support, nous avons : $\mathrm{Supp}(\overline{X}) = 1 - \mathrm{Supp}(X) = 1 - 0,6 = 0,4$. Dans cette situation, la génération d'une règle négative à partir des motifs positifs conduit à l'obtention d'une règle valide.

— **Exemple 2 :**
En prenant un $minSupp = 0,4$ et un motif X tel que $\text{Supp}(X) = 0,8$.



Le motif X est bien un motif fréquent. Par contre, $\overline{X}$ est loin d'être fréquent. En effet, $\text{Supp}(\overline{X}) = 1 - \text{Supp}(X) = 1 - 0,8 = 0,2$.

— **Exemple 3 :**
Cette fois, en prenant un $minsupp = 0,9$ et un motif X tel que $\text{Supp}(X) = 0.8$.
On tombe dans un cas où le motif X n'est même pas fréquent. Comme $1 - \text{Supp}(X)$

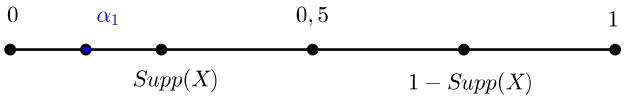


est plus petit que $\text{Supp}(X)$, le motif négatif $\overline{X}$ n'est pas fréquent non plus.

***Deuxième cas : Supp*$(X) < 0,5$**

Dans le présent cas, le support (X) est plus petit que celui de $\overline{X}$. À travers les trois exemples qui vont suivre, nous allons montrer que parmi les motifs positifs non fréquents, ceux qu'on n'a pas pris en considération dans les anciens algorithmes d'extraction des bases des règles valides selon la mesure M_{GK}, il y en a certains qui correspondent aux motifs négatifs fréquents.

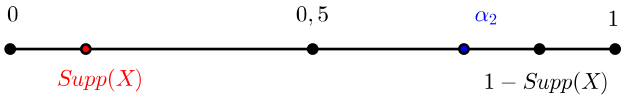
— **Exemple 4 :**
Prenons un $minsup = 0,3$ et un motif X tel que $\text{Supp}(X) = 0,4$.
Ici, X et $\overline{X}$ sont tous les deux fréquents $\text{supp}(\overline{X}) = 0,6$. Donc, on ne risque pas de



perdre les éventuelles règles constituées par ces deux motifs. Ce n'est pas le cas pour l'exemple 5.

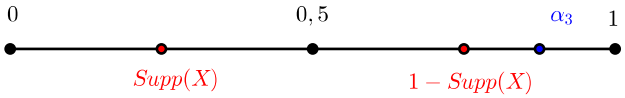
— **Exemple 5 :**
Prenons un $minSupp = 0,7$ et un motif X tel que $\text{Supp}(X) = 0,2$. Comme on peut le constater que le motif X est loin d'être fréquent. Par contre, $\overline{X}$ est un motif qui mérite d'être pris en considération $(\text{Supp}(\overline{X}) = 0,8)$. Encore une fois, on tombe dans un cas où la non considération d'un motif non fréquent, nous amène à la perte d'une règle négative qui pourrait être valide.

— **Exemple 6 :**
Dans ce dernier cas, on a fait en sorte que les motifs X et $\overline{X}$ soient tous les deux non



fréquents.

En prenant un support minimum très élevé ($minSupp = 0,9$) et un motif X tel que



$\text{Supp}(X) = 0,2$, on tombe dans ce dernier cas (X et $\overline{X}$ ne sont pas fréquents). L'énumération de ces différents cas possibles nous montre que pour un motif quelconque X (fréquent ou non), le motif $\overline{X}$ associé peut être fréquent comme il peut être non fréquent. En écartant donc ces motifs non fréquents des données à l'entrée des algorithmes d'extraction des bases des règles M_{GK}-valides, on risque de perdre des règles négatives valides. Pour éviter ce type de problème, dans nos algorithmes d'extraction des nouvelles bases, nous n'allons pas nous restreindre aux motifs fermés fréquents, mais nous allons considérer l'ensemble des motifs fermés de support non nul comme données de départ. Nous supprimons ensuite les motifs non fréquents au fur et à mesure.

5.4 Bases des règles négatives

Pour tous motifs X, Y d'un certain contexte d'extraction $\mathcal{K}$, si X défavorise Y, le degré ou l'intensité de répulsion de ces deux motifs peut constituer des informations très utiles. En effet, si on arrive à conclure que la présence des items constituant le motif X conduit systématiquement à l'absence de ceux qui constituent le motif Y (si X, alors $\overline{Y}$), l'utilisateur peut prendre des décisions adéquates relatives à ces deux motifs. Il est donc très intéressant d'étudier aussi s'il y a des liens négatifs dans nos données. Nous avons déjà donné des description des bases des règles négatives (règles de type $X \rightarrow \overline{Y}$). Certes, si X défavorise Y, alors X favorise $\overline{Y}$, mais si Y est fréquent, rien ne nous garantit de la représentativité de $\overline{Y}$ par rapport à un *minSupp*. Nous avons étudié les différents scénarios possibles au paragraphe 5.3. En tenant compte de toutes ces particularités, nous allons donner les algorithmes d'extraction des bases des règles négatives.

5.4.1 Bases des règles négatives exactes

Rappelons qu'une règle r est négative exacte lorsque qu'au moins un de ses composants (prémisse ou conséquent) est un motif négatif et $\text{M}_{GK}(r) = 1$. Ici, nous nous intéressons aux règles de type $X \rightarrow \overline{Y}$ telle que $\text{M}_{GK}(X \rightarrow \overline{Y}) = 1$. Nous avons donné au chapitre 4 (§ 4.5) des descriptions de la base des règles négatives exactes. Nous avons proposé l'ensemble $NBNE$ que nous allons rappeler ci-après :

$$NBNE = \left\{ \begin{array}{l} G_X \rightarrow \{\overline{x}\} : \ X \in Bd^+(0), G_X \in \mathcal{G}_X, \ \ x \notin X \\ et \ \min\left(Supp(X), \ Supp(\{\overline{x}\})\right) \geq minSupp \end{array} \right\}.$$

$\mathcal{G}_X$ désigne l'ensemble des générateurs de X. À partir de la connaissance de $Bd^+(0)$ et de l'ensemble des générateurs $\mathcal{G}_X$ des éléments X de $Bd^+(0)$ d'un contexte, nous pouvons proposer un algorithme de génération de $NBNE$ d'un quelconque contexte $\mathcal{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$.

Algorithme 7 Base des Règles Négatives Exactes

Entrée : $LBd^+(0)$: Liste contenant les éléments de $Bd^+(0)$ qui sont fréquents
Ensemble des générateurs des éléments X de $LBd^+(0)$ et le $minSupp$
Sortie : $NBNE$ Nouvelle Base Négative Exacte

1: $NBNE = \{\}$
2: **Pour** Chaque X dans $LBd^+(0)$ **faire**
3: $\quad\mathcal{G}_X = generator(X)$
4: $\quad$**Pour** Chaque G_X de $\mathcal{G}_X$ **faire**
5: $\quad\quad$**Pour** Chaque y de $\mathcal{I}$ **faire**
6: $\quad\quad\quad P_{yb} = 1 - Supp(\{y\})$ /* $P(\overline{\{y\}}) = 1 - P(\{y\})$ */
7: $\quad\quad\quad$**Si** ($P_{yb} \geqslant minSupp$ ET $y \notin X$) **alors**
8: $\quad\quad\quad\quad BNE = BNE \cup \{g_X \rightarrow \overline{\{y\}}\}$
9: $\quad\quad\quad$**Fin Si**
10: $\quad\quad$**Fin Pour**
11: $\quad$**Fin Pour**
12: **Fin Pour**
13: Retourner $NBNE$

L'algorithme 7 parcourt tous les $Bd^+(0)$ qui sont fréquents, c'est à dire, les éléments de $LBd^+(0)$ (ligne 2). Étant donné qu'une bordure positive est nécessairement un fermé et que ce dernier peut avoir un ou plusieurs générateurs, nous devons avoir à notre disposition l'ensemble des générateurs d'un motif X de $LBd^+(0)$: c'est le rôle de la fonction $generator()$ (ligne 3). Pour chacun de ces générateurs (qui sont nécessairement fréquents d'ailleurs, puisque le motif X l'est), on parcourt les items i de $\mathcal{I}$ pour trouver ceux qui sont fréquents et qui ne font pas partie de X (ligne 5 à 8). Chacun de ces items i va être le conséquent des règles exactes et valides de prémisse g_X.
Prenons l'exemple du contexte $\mathcal{K}$ décrit dans le tableau 2.1. Dans cet exemple, les bordures positives sont des motifs de support 1/6 (par rapport à 6 objets ou transactions). Pour que l'on puisse dérouler l'algorithme, on va devoir prendre un $minSupp = 1/6$.

$$BD^+(0) = LBD^+(0) = \{ABCEF,\ ABDEF,\ ABCD\}$$

Par rapport au $minSupp$ que nous venons de choisir, tous les items i de $\mathcal{I}$ sont fréquents, le test de l'algorithme 7 est résumé dans les tableaux ci-après.

$X = ABCEF$

G_X	Items $\mathcal{I}$ ($\{y\}$)	$Supp(\{\overline{Y}\})$	Règles négatives
CF	A	1/6	$CF \rightarrow \overline{D}$
	B	1/6	
	C	1/2	
	D	1/3	
	E	1/3	
	F	2/3	

$X = ABCD$

G_X	Items $\mathcal{I}$ ($\{y\}$)	$Supp(\{\overline{Y}\})$	Règles négatives
CD	A	1/6	$CD \rightarrow \overline{F}$
	B	1/6	
	C	1/2	
	D	1/3	
	E	1/3	$CD \rightarrow \overline{E}$
	F	2/3	

$X = ABDEF$

G_X	Items $\mathcal{I}$ ($\{y\}$)	$Supp(\{\overline{Y}\})$	Règles négatives
DF	A	1/6	$DF \rightarrow \overline{C}$
	B	1/6	
	C	1/2	
	D	1/3	
	E	1/3	
	F	2/3	

$X = ABDEF$

G_X	Items $\mathcal{I}$ ($\{y\}$)	$Supp(\{\overline{Y}\})$	Règles négatives
ADE	A	1/6	$ADE \rightarrow \overline{C}$
	B	1/6	
	C	1/2	
	D	1/3	
	E	1/3	
	F	2/3	

Tableau 5.5 – Exemple d'exécution de l'algorithme 7

À partir de ces tableaux, on peut avoir l'ensemble $NBNE$ du contexte donné dans le tableau 2.1 : $NBNE = \{CF \rightarrow \overline{D},\ CD \rightarrow \overline{F},\ CD \rightarrow \overline{E},\ DF \rightarrow \overline{C},\ ADE \rightarrow \overline{C}\}$.
Après avoir décrit un algorithme d'extraction d'une des règles négatives exactes, nous allons passer à la génération des bases des règles négatives approximatives.

5.4.2 Bases des règles négatives approximatives

Selon la définition 4.12, les éléments de la Nouvelle Base des règles Négatives Approximatives ($NBNA$) sont formés à partir des motifs générateurs. Dans la conception d'un algorithme d'extraction de $NBNA$, nous allons supposer que ces motifs générateurs sont disponibles. De plus, comme dans les autres bases des règles négatives, pour ne pas perdre des motifs négatifs fréquents et ne pas avoir des motifs négatifs non fréquents (pourtant les motifs positifs associés sont fréquents), nous allons considérer tous les générateurs de supports non nuls au lieu de se contenter des motifs positifs fréquents comme dans la plupart des algorithmes d'extraction des bases des règles. Ensuite, on sélectionne les prémisses dans l'ensemble des générateurs positifs fréquents et les conséquents dans l'ensemble des générateurs négatifs fréquents. Quant à la mesure M_{GK}, nous avons vu dans les propriétés de cette mesure que lorsqu'on a deux motifs X et Y tels que X défavorise Y (i. e. $P(Y'/X') - P(Y') < 0$), on a l'égalité : $M^f_{GK}(X \rightarrow \overline{Y}) = -M^d_{GK}(X \rightarrow Y)$. Rappelons que :

$$\begin{cases} M^f_{GK}\left(X \rightarrow \overline{Y}\right) = \frac{P\left(\overline{Y}'/X'\right) - P\left(\overline{Y}'\right)}{1 - P\left(\overline{Y}'\right)}, \\ M^d_{GK}\left(X \rightarrow Y\right) = \frac{P(Y'/X') - P(Y')}{P(Y')}. \end{cases}$$

Nous allons donc nous servir de cette propriété pour avoir les mesures des règles négatives

en ne manipulant que des motifs positifs.
Avant de donner un algorithme d'extraction de $NBNA$, voici un tableau (cf. Tableau 5.6 qui résume les principales variables utilisées.

LG	Liste des générateurs de support non nul
CP	Liste des Candidats prémisses
CC	Liste des Candidats conséquents

Tableau 5.6 – Quelques variables utilisées dans l'algorithme 8

Algorithme 8 Base des Règles Négatives Approximatives

Entrée : Ensemble des générateurs de support non nul et le $minSupp$
α : Niveau de confiance utilisé pour le calcul de valeur théorique de χ^2
Sortie : $NBNA$ Nouvelle Base Négative Approximative

1: $Khi2 = Khi2(\alpha), CP = \{\}, CC = \{\}, NBNA = \{\}$
2: **Pour** Chaque X dans LG **faire**
3: **Si** $(Supp(X) \geqslant minSupp)$ **alors**
4: $CP = CP \cup X$
5: **Fin Si**
6: **Si** $(1 - Supp(X) \geqslant minSupp)$ **alors**
7: $CC = CC \cup X$
8: **Fin Si**
9: **Fin Pour**
10: **Pour** Chaque $X \in CC$ **faire**
11: **Pour** Chaque $Y \in CP$ **faire**
12: **Si** $(Y \neq X)$ **alors**
13: $n_X = Card(X')$
14: $n_Y = Card(Y')$; $n_{XY} = Card(X' \cap Y')$
15: $Ecart = n_{XY}/n_X - n_Y/n$
16: **Si** $Ecart < 0$ **alors**
17: mgk $= -Ecart/n_Y/n)$
18: mgkcr $= \sqrt{\frac{1}{n}\frac{n-n_X}{n_X}\frac{n-n_Y}{n_Y}Khi2}$
19: **Si** mgk $\geq$ mgkcr **alors**
20: $NBNA = NBNA \cup \{X \rightarrow \overline{Y}\}$
21: **Fin Si**
22: **Fin Si**
23: **Fin Si**
24: **Fin Pour**
25: **Fin Pour**
26: Retourner $NBNA$

Dans cet algorithme, on commence par construire l'ensemble CP, ensemble des motifs susceptibles d'être la prémisse d'une règle, et l'ensemble CC des motifs qui peuvent être des conséquents des règles. Ces ensembles sont générés à partir des générateurs de support non nul (ligne 2 à 9). Une fois que ces deux ensembles sont construits, il suffit de les croiser pour former, tester et éventuellement pour valider des règles de type $X \rightarrow \overline{Y}$. Une fois que la mesure M_{GK} d'une règle $X \rightarrow \overline{Y}$ dépasse sa valeur critique calculée au niveau de confiance α, on récupère la règle et on la met dans $NBNA$ (ligne 10 à 20).

5.4.3 Semi-base des règles négatives approximatives

Comme dans les autres semi-bases, pour générer la semi-base négative-approximative, nous pouvons partir de la $NBNA$ et comparer deux à deux les règles qui constituent cet ensemble en utilisant la relation d'ordre « dominée par » définie au paragraphe 4.8 concernant la relation d'ordre dans l'ensemble des règles négatives. Rappelons qu'une règle $r_2 : g_{X_2} \rightarrow \overline{g_{Y_2}}$ est dominée par une autre règle $r_1 : g_{X_1} \rightarrow \overline{g_{Y_1}}$ si et seulement si :

$$\begin{cases} r_1 \, et \, r_2 \;\; sont \; valides, \\ g_{X_1} \subseteq g_{X_2} \; et \; g_{Y_1} \subseteq g_{Y_2}. \end{cases}$$

Nous allons décrire un algorithme permettant de comparer deux règles. Pour toutes règles r_1, r_2 de l'ensemble des règles négatives approximatives, l'algorithme fournira celle qui est dominante ou l'ensemble vide lorsque les deux règles ne sont pas comparables.

Algorithme 9 Fonction CompNE

Entrée : r_1, r_2 deux règles négatives
Sortie : r règles dominantes ou $\emptyset$

1: **Si** (r_1.Premisse $\subseteq$ r_2.Premisse ET r_1.Consequent $\subseteq$ r_2.Consequent) **alors**
2: $r = r_1$
3: **Sinon**
4: **Si** (r_2.Premisse $\subseteq$ r_1.Premisse ET r_2.Consequent $\subseteq$ r_1.Consequent) **alors**
5: $r = r_2$
6: **Sinon**
7: $r = \emptyset$
8: **Fin Si**
9: **Fin Si**
10: Retourner r

Possédant la fonction CompNE(), nous pouvons comparer les éléments de $NBNA$ et supprimer les règles qui sont dominées par d'autres. L'algorithme de génération de $SBNA$ sera très semblable à l'algorithme 5. Les différences se trouvent juste au niveau de l'entrée de l'algorithme et à la fonction de comparaison des règles utilisée.
On prend un à un les éléments de $NBNA$ et on compare chaque règle avec les restes des éléments de $NBNA$. La variable EspaceTest contient les éléments qui peuvent être comparés à une règle fixée au départ. Si on a trouvé un élément qui domine la règle de départ dans l'EspaceTest, alors la règle sera écartée de SBNA et de son EspaceTest doit être réduit à l'ensemble vide pour arrêter la boucle et recommencer avec une autre règle de $NBNA$. Dans le cas contraire, il faut comparer la règle de départ à tous les éléments de son EspaceTest

Algorithme 10 Sémi-Base des Règles Négatives Approximatives

Entrée : $NBNA$
Sortie : $SBNA$
1: $SBNA = NBNA$
2: **Pour** Chaque r_1 de $NBNA$ **faire**
3: EspaceTest = $SBNA \setminus r_1$
4: **Pour** Chaque r_2 dans EspaceTest **faire**
5: r =CompNE(r_1, r_2)
6: **Si** $(r == r_1)$ **alors**
7: $SBNA = SBNA \setminus r_2$
8: **Fin Si**
9: **Si** $(r == r_2)$ **alors**
10: $SBNA = SBNA \setminus r_1$; EspaceTest = $\emptyset$
11: **Fin Si**
12: **Fin Pour**
13: **Fin Pour**
14: Retourner $SBNA$

(ligne 4 à 12). Dans le cas où les règles ne sont pas comparables, on ne fait rien. C'est ainsi qu'on élimine petit à petit les règles dominées par d'autres dans le processus de construction de $SBNA$.

5.5 Conclusion partielle

Ce chapitre nous a permis de décrire quelques algorithmes d'extraction des nouvelles bases des règles. À travers les quelques exemples d'exécution de ces algorithmes, nous avons pu apprécier les différences entre les anciennes et les nouvelles bases des règles M_{GK}-valides. En plus des algorithmes d'extraction des bases, nous avons proposé aussi des algorithmes d'extraction des semi-bases. Ils sont basés sur la comparaison des règles valides. Autrement dit, les algorithmes d'extractions des semi-bases sont conçus à partir de la relation d'ordre partiel définie dans chacune des catégories des règles.
À l'entrée de ces algorithmes, nous avons besoin des motifs fermés, des générateurs et des bordures positives. Ces besoins peuvent être perçus comme limite de ces algorithmes. D'un autre côté, par rapport aux algorithmes performants qui sont disponibles dans la littérature, on peut citer entre autres, les algorithmes d'extraction des motifs fermés (Close, A-Close, Close$^+$), extraction des motifs maximaux (MaxClique, MaxEclat, MaxMiner), extraction des générateurs (JEN), nous avons jugé stratégique de pouvoir compter sur ces algorithmes d'extraction des motifs particuliers. En fait, l'idée des algorithmes présentés dans ce chapitre est de montrer comment manipuler ces motifs particuliers pour avoir les nouvelles bases, y compris les semi-bases.
Après la présentation de ces algorithmes, il est tout à fait normal de se lancer à l'implémentation. Nous laissons cette partie des taches en perspectives. Pour apprécier concrètement la qualité des règles qui peuvent se trouver dans ces nouvelles bases M_{GK}-valides, nous allons utiliser dans le prochain chapitre les théories que nous avons présentées pour étudier les données réelles issues de nos expérimentations didactiques.

Chapitre 6

Application des bases des règles M_{GK}-valides dans une expérimentation

Sommaire

6.1 Introduction

Après avoir fait des descriptions et donné des algorithmes d'extraction des bases et semi-bases des règles valides selon la mesure M_{GK}, nous allons utiliser ces concepts pour analyser des données concrètes. Dans ce chapitre, nous allons essayer d'atteindre les deux objectifs suivants :

1. Découverte des informations cachées dans des données concrètes en se servant du couple des mesures $Support - \mathrm{M}_{GK}$,
2. Études comparatives des nouvelles bases et semi-bases des règles.

Quand on analyse un contexte concret, on doit s'attendre à un volume important de données. Ce qui nécessite un traitement automatisé, pourtant, dans notre cas, nous n'avons pas encore implémenté les nouveaux algorithmes d'extraction des nouvelles bases et semi-bases. Néanmoins, pour ce cas actuel des données, nous nous contentons d'un traitement ad hoc à titre prototypique. Nous sommes donc contraints à combiner plusieurs applications pour effectuer l'extraction des bases des règles M_{GK}-valides. Nous avons utilisé le « package arules du logiciel R et le logiciel Tanagra » pour les extractions des motifs particuliers (fréquents, Fermés, Générateurs) ; ensuite, nous avons aussi utilisé « OpenOffice calc » pour les calculs de valeurs critiques de M_{GK} et les sélections des règles valides. Cette approche nous a contraint à limiter le nombre des motifs à étudier. Toutefois, nous allons faire des choix qui nous permettront d'atteindre les deux objectifs cités ci-dessus. Nous commencerons par extraire des bases, des semi-bases ou encore, quelques éléments constituants les bases. Cette phase

sera suivi des interprétations sur les connaissances extraites et sur la forme et le nombre des éléments constituants les bases.

6.2 Base et semi-base des règles positives exactes

Nous allons reprendre les données de l'expérimentation présentée au chapitre 2 (§ 2.5) concernant l'étude comparative des approches pédagogiques avec ou sans assistance des TICE. Si nous nous sommes servis de la mesure « intensité d'implication » à travers l'utilisation du logiciel CHIC pour analyser ces données, maintenant nous allons refaire les analyses (avec les mêmes données) mais cette fois, avec le couple des mesures $Support - \mathrm{M}_{GK}$ en générant des bases ou quelques éléments constituant les bases des règles $Support - \mathrm{M}_{GK}$-valides.

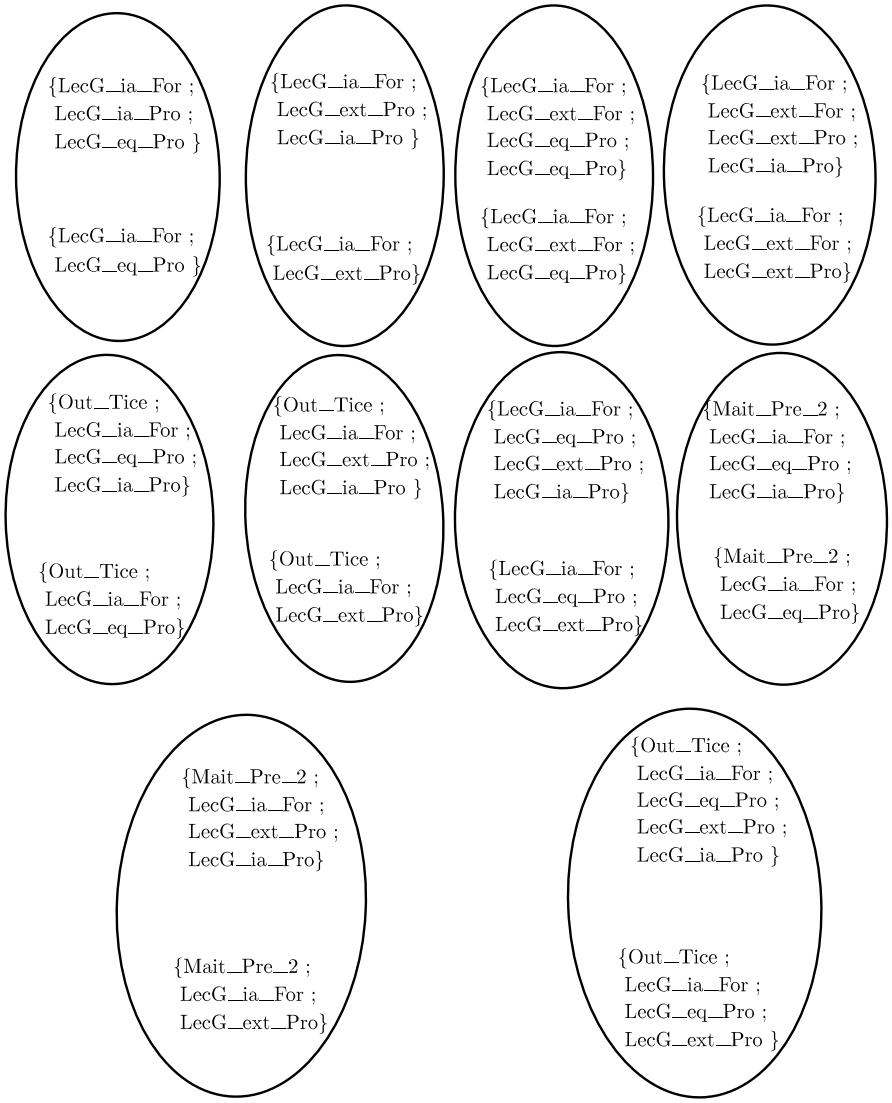


FIGURE 6.1 – Classes des motifs de même fermeture ayant plus d'un élément

En fixant un $minSup = 0,33$, nous arrivons à 88 motifs fréquents dont 78 sont fermés. Dans ce contexte, il n'y a ainsi que 10 motifs fréquents qui ne sont pas fermés. Les classes contenant ces motifs sont représentées par la figure 6.1. Les autres classes sont constituées par un

seul motif (donc un fermé). Rappelons que les règles positives exactes sont générées à partir des motifs (prémisses et conséquents) qui se trouvent dans une même classe de fermeture. Évidemment, on ne peut pas générer des règles exactes avec une classe qui ne contient qu'un seul motif. À partir de ces 10 classes, on peut ainsi générer 10 règles qui vont constituer la base des règles positives exactes qui sont consignées dans le Tableau 6.1.

I	r_1	LecG_ia_For - LecG_ext_Pro - Out_Tice s - LecG_eq_Pro	==>	LecG_ia_Pro
	r_2	LecG_ia_For - Mait_Pre_2 - LecG_eq_Pro	==>	LecG_ia_Pro
	r_3	LecG_ia_For - LecG_eq_Pro - LecG_ext_For	==>	LecG_ia_Pro
	r_4	LecG_ia_For - Out_Tice s - LecG_eq_Pro	==>	LecG_ia_Pro
	r_5	LecG_ia_For - LecG_eq_Pro	==>	LecG_ia_Pro
II	r_6	LecG_ia_For - Mait_Pre_2 - LecG_ext_Pro	==>	LecG_ia_Pro
	r_7	LecG_ia_For - LecG_ext_Pro - LecG_eq_Pro	==>	LecG_ia_Pro
	r_8	LecG_ia_For - LecG_ext_Pro - Out_Tice	==>	LecG_ia_Pro
	r_9	LecG_ia_For - LecG_ext_Pro	==>	LecG_ia_Pro
III	r_{10}	LecG_ia_For - LecG_ext_Pro - LecG_ext_For	==>	LecG_ia_Pro

Tableau 6.1 – Base Positive Exacte BPE

Passons maintenant à l'interprétation de ces résultats. Examinons les règles dans cette base positive exacte. Puisque les classes à partir desquelles on a construit BPE ne contenaient que deux éléments de chaque, l'ensemble des règles exactes valides et les éléments de BPE coïncident. Autrement dit, ici, l'extraction de BPE n'apporte pas de réduction des règles exactes valides à interpréter. Nous pouvons remarquer que pour ces 10 règles, on a toujours le même conséquent LecG_ia_Pro ; c'est-à-dire que la réalisation de l'un des prémisses de ces 10 règles implique la réalisation du motif « LecG_ia_Pro ».
Si nous prenions une à une ces 10 règles. Selon r_1 par exemple, dans ce contexte d'expérimentation, les élèves qui :

— savent lire graphiquement l'image ou un antécédent dans une situation formelle [1] sans modélisation, pas d'interprétation, pas d'application (LecG_ia_For),

— arrivent à lire graphiquement l'extremum (minimum ou maximum) d'une fonction dans une situation problème [2] (LecG_Ext_Pro),

— savent comment résoudre graphiquement une équation donnée dans une situation problème (LecG_eq_Pro),

— ont pu bénéficier les aspects dynamiques des activités vidéo-projetés en supplément de l'utilisation d'un tableau noir (Out_Tice s),

ont certainement réussi à effectuer la lecture graphique de l'image ou d'un antécédent dans une situation problème (LecG_ia_Pro), la règle r_1 étant une règle exacte.

1. Situations auxquelles on donne un exercice avec des expressions formelles, (§ 2.5.4).
2. Ici on donne les exercices sous forme de problèmes plus ou moins concrets (§ 2.5.4).

Exploitations possibles de ces informations

Nous allons voir comment exploiter ces types d'informations (les interprétation de ces 10 règles) dans un exercice d'enseignement. Pour cela, nous allons examiner l'utilité de la règle r_2. Les informations véhiculées par la règle r_2 peuvent être utilisées pour ordonner les difficultés susceptibles d'être rencontrés par les élèves lors d'un processus d'apprentissage. Prenons, par exemple, une situation où un enseignant devait optimiser le nombre de questions (nombre d'exercices) à poser à ses élèves. Selon la règle exacte r_2, un élève qui sait lire graphiquement l'image ou un antécédent dans une situation formelle (LecG_ia_For), qui maitrise les prérequis sur les équations de second degré (Mait_Pre_2) et qui arrive à résoudre graphiquement une équation posée dans une situation problème (LecG_eq_Pro) sera certainement (r_2 est une règle exacte) capable d'effectuer la lecture graphique de l'image ou d'un antécédent dans une situation problème (LecG_ia_Pro). Donc, pour une raison ou une autre, si un enseignant est « certain » que ses élèves peuvent réussir les trois items (LecG_ia_For, Mait_Pre_2 et LecG_eq_Pro), ce n'est plus nécessaire de leur proposer des exercices qui visent à tester s'ils (les élèves) savent lire graphiquement l'image ou un antécédent dans une situation problème. Ces règles peuvent être encore plus intéressantes si on les interprète en prenant les contraposées. Nous avons montré au chapitre 2 (sur la raison du choix de mesure M_{GK}) que la compatibilité avec la logique classique est l'un des points forts da la mesure M_{GK}. En effet, selon cette mesure, dans le cas d'attraction mutuelle de deux motifs X et Y et sous réserve de le représentativité des motifs X, Y, $\overline{X}$ et $\overline{Y}$, on a une équivalence entre la validité de la règle $X \to Y$ et celle de $\overline{Y} \to \overline{X}$ (puisque $M^f_{Gk}(X \to Y) = M^f_{Gk}(\overline{Y} \to \overline{X})$).

La contraposée de la règle r_3 nous affirme que si un élève ne réussit pas à faire la lecture graphique de l'image ou d'un antécédent dans une situation problème, alors il ne peut pas réussir l'un ou l'autre ou une combinaison des trois items LecG_ia_For, LecG_eq_Pro et LecG_ext_For. Autrement dit, pour aider un élève qui a un problème sur l'un ou l'autre ou la combinaison des trois items constituant la prémisse de la règle r_3, il faut d'abord insister sur l'item concernant la lecture graphique de l'image ou d'un antécédent dans une situation problème (LecG_ia_Pro). Ce type d'information est très utile surtout pour un tuteur artificiel ou un enseignement non presentiel. Tant que les problèmes sur la lecture graphique de l'image ou d'un antécédent ne sont pas résolus, il ne faut pas passer aux trois items LecG_ia_For, LecG_eq_Pro et LecG_ext_For. Donc, r_3 nous donne une taxonomie des thèmes d'enseignements à traiter. Conformément à certains résultats fournis par CHIC, l'item qui identifie la capacité d'effectuer une lecture graphique de l'image et d'un antécédent dans une situation problème se trouve au niveau « facile » si on le compare aux trois items : LecG_ia_For - LecG_eq_Pro - LecG_ext_For. Puisque la lecture graphique de l'image et d'un antécédent semble plus « facile » à l'égard des élèves, pour aborder les thèmes identifiés par la prémisse de r_3, l'enseignant a tout intérêt de commencer par traiter le thème identifié par l'item LecG_ia_Pro. Ce résultat justifie aussi l'idée selon laquelle il est important de commencer l'enseignement d'un thème fixé par des activités plus ou moins concrètes, des activités liées à des situations connues par les élèves avant de passer aux activités qui nécessitent de l'abstraction et des calculs formels. Soulignons aussi que le passage à la contraposée d'une règle positive, en l'occurrence la règle r_3, nous permet de justifier, encore une fois, le fait de choisir un motif de taille minimale pour le conséquent d'un élément d'une base des règles négatives à droite. En effet, selon l'interprétation que nous venons de faire, le fait d'échouer à l'item LecG_ia_Pr implique l'échec à l'un ou à l'autre ou à la combinaison des trois items LecG_ia_For, LecG_eq_Pro et LecG_ext_For. Cette information se précise de plus en plus au

fur et à mesure de la diminution de la taille du motif négatif conséquent. On peut faire le même raisonnement pour les prémisses négatifs. La précision d'information apportée par une règle négative croit aussi en fonction de la diminution de la taille des motifs négatifs intervenant dans les règles. Nous pouvons remarquer que certaines prémisses de ces 10 règles sont comparables et, si l'on devait trouver les conditions suffisantes pour que ses élèves réussissent la lecture graphique de l'image ou d'un antécédent dans une situation problème, on chercherait à atteindre son objectif avec les conditions les moins contraignantes. Cette affirmation nous amène à la notion de semi-base des règles.

Semi-base positive exacte
Selon r_1, réussir simultanément les quatre thèmes d'enseignements identifiés par les items LecG_ia_For, LecG_ext_Pro, Out_Tic s et LecG_eq_Pro suffisent pour que l'on puisse affirmer que l'élève réussira aussi la lecture graphique de l'image et d'un antécédent dans une situation problème (LecG_ia_Pro). Selon r_2, il faut seulement réussir simultanément les trois items LecG_ext_Pro, Out_Tice s et LecG_eq_Pro. Pour avoir une même conséquence, la condition suffisante posée par r_2 est moins contraignante que celle posée par r_1, et ainsi de suite. Nous pouvons donc remarquer que toutes les règles de la ligne I du tableau 6.1 sont comparables selon la relation d'ordre définie dans l'ensemble des règles positives exactes (§ 4.8.1). On peut faire la même remarque pour les lignes II et III. D'où la semi-base positive exacte de notre contexte d'expérimentation présentée dans le tableau 6.2 ci-dessous :

LecG_ia_For - LecG_eq_Pro	==>	LecG_ia_Pro
LecG_ia_For - LecG_ext_Pro	==>	LecG_ia_Pro
LecG_ia_For - LecG_ext_Pro - LecG_ext_For	==>	LecG_ia_Pro

Tableau 6.2 – Semi-base positive exacte ($SBPE$)

En fait, au lieu d'interpréter r_1 par exemple, on peut se contenter de r_5 en affirmant que les élèves qui arrivent à lire graphiquement et correctement l'image et un antécédent dans une situation formelle (LecG_ia_For) et la lecture graphique des solutions d'équation dans une situation problème (LecG_eq_Pro) sont toujours capables de lire graphiquement l'image et un antécédent dans une situation problème (LecG_ia_Pro). On peut reprendre le même raisonnement utilisé dans l'interprétation des règles dans BPE pour interpréter les règles dans $SBPE$. La semi-base fournit des résultats beaucoup plus compacts que ceux de BPE.

6.3 Base et semi-base des règles approximatives

Dans notre contexte, pour trouver tous les éléments qui composent la nouvelle base positive approximative avec le support minimum $minSupp = 0.33$, il faut croiser les 78 motifs fermés fréquents. Étant donné le nombre de tests qu'on doit effectuer, nous ne pouvons pas générer la base positive approximative sans un programme approprié. Pour cela, nous allons interpréter quelques règles constituant la nouvelle base positive approximative ($NBPA$). Le tableau 6.3 regroupe les éléments de $NBPA$, générés au niveau de confiance $\alpha = 0,98$. Rappelons que l'item Out_Tice.s désigne une approche pédagogique basée sur l'utilisation simultanée et appropriée d'un tableau noir et d'un vidéoprojecteur.

Règles (r_i)	Prémisses : X	Conséquents : Y	$P(X)$	$P(Y)$	$M_{GK}(r_i)$	$M_{GK}^{(0,98)}(r_i)$
r_1	Out_Tice.s-Mait_Pre_2	Df_Calc_For	0, 51	0, 37	0, 23	0, 16
r_2	Out_Tice.s-Mait_Pre_2	Util_Def_F_autr	0, 51	0, 43	0, 26	0, 18
r_3	Out_Tice.s-Mait_Pre_2	Util_Def_F_autr - Df_Calc_For	0.51	0, 33	0, 20	0, 14
r_4	Out_Tice.s-Mait_Pre_2	Df_Calc_For-LecG_ia_Pro	0, 51	0, 35	0, 21	0, 15
r_5	Out_Tice.s-Mait_Pre_2	Util_Def_F_autr-LecG_ia_For	0, 51	0, 33	0, 18	0, 14
r_6	Out_Tice.s-Mait_Pre_2	Util_Def_F_autr-LecG_ia_Pro	0, 51	0, 39	0, 21	0, 16
r_7	LecG_ine_For	LecG_eq_For	0, 12	0, 22	0, 72	0, 31
r_8	LecG_eq_For	LecG_ext_For	0, 22	0, 60	0, 63	0, 48
r_9	LecG_eq_Pro	LecG_ia_Pro-LecG_ext_Pro-LeG_ia_For	0, 63	0, 54	0, 23	0, 18
r_{10}	LecG_eq_Pro	LecG_ia_Pro-LecG_ext_Pro	0, 63	0, 67	0, 28	0, 23
r_{11}	LecG_eq_Pro	LecG_ia_Pro-LeG_ia_For	0, 63	0, 69	0, 35	0, 25
r_{12}	LecG_eq_Pro	LecG_ia_Pro	0, 63	0, 90	0, 87	0, 49
r_{13}	LecG_eq_Pro	LecG_ia_Pro-LecG_ext_Pro-LeG_ia_For-LecG_ext_For	0, 63	0, 35	0, 13	0, 12
r_{14}	Util_Def_F_autr	Mait_Pre_2	0, 43	0, 73	0, 72	0, 40
r_{15}	Util_Def_F_autr	Mait_Pre_2-Df_Calc_For	0, 43	0, 35	0, 59	0, 18
r_{16}	Util_Def_F_autr	Out_Tice.s-Mait_Pre_2	0, 43	0, 51	0, 37	0, 25

Tableau 6.3 – Base Positive Approximative BPA

Selon les règles r_1 à r_6, au niveau de confiance 98% (risque d'erreur de 2%), les élèves qui maîtrisent les prérequis relatifs aux équations de second degré (Mait_Pre_2) et qui ont reçu des enseignements d'un professeur exploitant simultanément un tableau noir et un vidéoprojecteur sont capables de (d') :

- r_1 : réussir la recherche de l'ensemble de définition dans une situation formelle,
- r_2 : utiliser la définition d'un ensemble de définition des fonctions non rationnelles,
- r_3 : réussir la recherche de l'ensemble de définition (d'une fonction rationnelle) et d'utiliser la définition de l'ensemble de définition des fonctions non rationnelles,
- r_4 : réussir la recherche de l'ensemble de définition et de lire graphiquement l'image et un antécédent dans une situation problème,

r_5 : utiliser la définition de l'ensemble de définition des fonctions non rationnelles et de lire graphiquement l'image et un antécédent dans une situation formelle,

r_6 : utiliser la définition de l'ensemble de définition des fonctions non rationnelles et de lire graphiquement l'image et un antécédent dans une situation problème.

Nous avons choisi d'interpréter ces 6 règles pour montrer certaines redondances. On peut par exemple, comparer les informations valides au niveau de confiance $\alpha = 0,98$ et apporter par les deux règles r_1 et r_4 (resp. r_2 et r_5). Le conséquent de r_1 (resp. de r_2) est une partie du conséquent de r_4 (resp. r_5) et les deux règles r_1 et r_4 (resp. r_2 et r_5) ont la même prémisse. Donc, connaissant les informations apportées par r_4 (resp. r_5), il n'est plus nécessaire (du moins en première approche) de présenter et interpréter la règle r_1 (resp. r_2). Par rapport à ces 6 règles, on peut générer des éléments de semi-base positive approximative au niveau de confiance $\alpha = 0,98$ en supprimant les deux règles r_1 et r_2. On voit que l'utilisation de semi-base permet de réduire le nombre des règles à interpréter en laissant à côté celles qui apportent des informations moins générales que d'autres. Le même raisonnement peut s'appliquer à l'ensemble des règles r_9 à r_{13}. Les conséquents de r_9 à r_{12} sont tous des parties du conséquent de r_{13} et toutes ces règles sont valides au niveau de confiance $\alpha = 0,98$. Donc, les informations apportées par les règles r_9 à r_{12} sont moins générales que celle apportée par r_{13}. Par conséquent, par rapport aux règles r_9 à r_{13}, c'est seulement r_{13} qui va constituer la semi-base positive approximative au niveau de confiance $\alpha = 0,98$. Intéressons-nous maintenant aux règles r_7 et r_8. Selon les résultats fournis par CHIC, ces deux règles sont valides au niveau de confiance $\alpha = 0,98$. Si on examine les motifs qui constituent ces deux règles, seul le motif LecG_ext_For est fréquent. Donc, avec un $minSupp = 0.33$, ces règles ne peuvent pas être valides. Bien que CHIC fournisse l'occurrence de chaque item du contexte d'extraction, il ne donne pas le support des itemsets (motifs) et il n'offre pas la possibilité de restreindre le graphe implicatif sur les motifs fréquents (par rapport à un *minSupp* fixé). Autrement dit, dans le graphe implicatif de ce contexte, on a un motif qui n'a que 12% de présence chez les individus (ici, les élèves). Si la mesure « intensité d'implication » valide ce type de règle, ce n'est pas le cas pour le couple de mesure « Support-M_{GK} ». On peut donc en conclure que le couple de mesures Support-M_{GK} est plus sélectif que l'intensité d'implication.

6.4 Conclusion partielle

L'extraction des bases des règles d'association est une méthode d'analyse des données permettant, entre autres, de découvrir des informations relatives aux liens de cause à effet dans une observation ou expérimentation didactique, notamment en didactique de mathématiques. En mettant de côté les règles moins informatives que d'autres, l'utilisation de la notion de semi-base permet de diminuer considérablement le nombre des règles à interpréter et cela, sans perdre des informations. Les règles valides permettent de créer une taxonomie des thèmes d'enseignement. Selon les données recueillies à l'issue de nos expérimentations et à la lumière de l'analyse de ces données, nous constatons que contrairement à ce qu'on rencontre en cours de mathématiques où l'on commence (souvent) les activités d'enseignement par des « définitions, théorèmes ou propriétés » pour éventuellement finir avec des « applications », il est plus judicieux d'inverser ces étapes. C'est à dire, avec l'assistance des outils informatiques, en exploitant l'aspect dynamique des ressources vidéoprojetées, il est préférable de commencer l'enseignement d'un thème précis par des activités plus ou moins concrètes avant d'enchainer sur les formulations abstraites.

Chapitre 7

Conclusions générales et perspectives

7.1 Utilisation des bases des règles dans les expérimentations

Dans la recherche en didactique, notamment en didactique des mathématiques, on a souvent besoin de recourir aux expérimentations, observations ou encore, aux évaluations et d'analyser les données recueillies pour valider ou rejeter des conjectures ou des hypothèses. Plusieurs outils statistiques peuvent être utilisés pour analyser les données recueillies. Dans la recherche en didactique, on peut avoir besoin de connaître s'il y des liens de cause à effet entre les observations. On peut citer, par exemple, l'importance de la découverte des liens entre les erreurs des élèves, entre les connaissances acquises, ou encore, entre les erreurs et les approches pédagogiques utilisées. D'où l'importance de l'utilisation de l'Analyse Statistique Implicative (ASI) pour les chercheurs en didactique des disciplines.
L'expérimentation que nous avons effectuée avait pour objectif d'établir des liens entre les connaissances que les élèves peuvent acquérir et l'approche pédagogique utilisée. On a voulu montrer les apports de l'utilisation des TICE sur le plan acquisition des connaissances. Certes, la taille de notre base des données est relativement petite. Malgré cela, nous pouvons tirer des résultats qui pourraient être généralisés.
Par rapport à nos objectifs, nous avons pu conclure que les élèves enseignés à la manière traditionnelle (activités donnés en utilisant le tableau et sans assistance des TIC) réussissent mieux les activités proposé en situation formelle ; par contre, ceux qui ont été enseignés avec l'outil tableau noir assisté par les TIC brillent sur les activités proposées en situation problème. Trois grandes lignes d'idées plus ou moins complémentaires peuvent être tirées de cette expérimentation. La première est la mise en garde à l'endroit des enseignants qui pensent que les diapositives peuvent remplacer les tableaux (noirs ou blancs). Dans l'enseignement de mathématiques, certaines activités, en l'occurrence les démonstrations des propriétés ou des théorèmes nécessitent des actions « en directe » et surtout pas « en différé » (préparer les diapositives et les projeter en classe). En effet, la plupart des élèves peuvent devenir très passifs lorsqu'on leur présente des concepts abstraits en diapositive. Par contre, au tableau, l'enseignant peut repérer facilement ceux qui décrochent en un instant précis et il pourrait interagir en conséquence (changer de discours ou de méthode, faire des rappels...) afin que la majorité de la classe reste active. D'autre part, cette expérimentation nous a permis de justifier des arguments qui pourraient être utilisés pour convaincre les enseignants qui s'accrochent (pour des tas des bonnes raisons) aux approches traditionnelles sans ordinateur

dans le processus d'enseignement et d'apprentissage. Nous avons pu montrer qu'une partie des objectifs de l'enseignement de mathématiques, à savoir la faculté d'utilisation des outils mathématiques pour résoudre des problèmes concrets peut être atteinte en se servant des outils informatiques. Si on constate actuellement que dans des nombreux systèmes éducatifs, l'enseignement des mathématiques reste très formel et n'a qu'un faible lien avec le monde réel, on peut contourner ces problèmes en mettant l'accent sur les simulations. De plus, les outils informatiques permettent de faire des conjectures et ces dernières sont des stimulants des réflexions car elles permettent de poser des questions adéquates. Donc, quel que soit les raisons évoquées, les enseignants ne doivent pas priver leurs élèves de ces merveilles technologiques.
Nous arrivons maintenant à la troisième grande ligne que nous avons pu tirer de notre expérimentation. En tenant compte des deux idées que nous venons d'expliciter, nous arrivons à la conclusion ci-après. Les enseignants doivent trouver un juste équilibre entre l'utilisation des technologies éducatives et celle d'un tableau. D'où la necéssité de la formation continue des enseignants, en particulier ceux qui n'ont pas été formés à l'utilisation de ces nouveaux outils. Prenons le cas du système éducatif Malagasy, si les dirigeants se sentent vraiment concernés par les problèmes de désaffection des jeunes à l'égard des disciplines scientifiques, en particulier les mathématiques, et leurs éventuelles conséquences, ils doivent promouvoir la « vraie » intégration pédagogique des TICE.

7.2 Bases des règles M_{GK}-valides

Dans un contexte d'extraction quelconque $\mathcal{K}$, il peut y avoir un nombre trop important de règles valides selon la mesure M_{GK} dont la plupart sont redondantes. Dans ce cas, découvrir objectivement des connaissances utiles dans l'ensemble de toutes ces règles valides est une tâche très difficile, voir impossible à réaliser ; d'où l'importance des concepts sur les bases des règles. De ce fait, puisque les bases sont les noyaux représentant l'ensemble des règles valides, elles doivent être constituées par les règles les plus informatives, c'est à dire les règles qui apportent les maximums d'informations. Dans la littérature, les chercheurs sont plus ou moins unanimes sur le fait qu'une règle positive de prémisse minimale et de conséquent maximal est plus informative que d'autres. Par rapport à cela, nous avons pu proposer quatre nouvelles bases des règles M_{GK}-valides : bases des règles positives et négatives exactes, bases des règles positives et négatives approximatives. Plusieurs points différencient les anciennes et ces nouvelles bases : comme les choix de prémisse et conséquent en tenant compte du fait que les règles les plus informatives ont des prémisses minimales et conséquents maximaux, la disjonction entre la prémisse et le conséquent d'une règle. Tout ceci a été réalisé grâce à l'introduction du concept des règles représentantes. Quant aux règles négatives, selon notre analyse et celle des autres chercheurs, une règle négative est plus informative, donc plus importante lorsque ses éléments constituants (prémisse et conséquent) sont minimaux. Nous avons conçu les bases des règles négatives (exactes et approximatives) en prenant des règles de prémisses et conséquents minimaux. Ainsi, les règles constituant les nouvelles bases des règles M_{GK}-valides sont nettement plus informatives que celles qui constituaient les anciennes bases. Par rapport au nombre des règles dans les bases, si on compare les nouvelles et les anciennes, nous n'avons pas pu réduire le nombre des règles dans les bases, mais nous avons réussi à améliorer la qualité des règles dans ces bases.

7.3 Semi-bases des règles M_{GK}-valides

Dans la littérature, on peut comparer deux ou plusieurs règles de même mesure ; celles qui ont de prémisses minimales et conséquents maximaux sont les plus informatives. Nous avons généralisé ce principe en comparant deux ou plusieurs règles de mesure qui ne sont pas forcément les mêmes. Nous avons utilisé l'hypothèse selon laquelle, pour un niveau de confiance α fixé par l'utilisateur, ce dernier est disposé à prendre en compte et au même pied toutes les informations (les règles) valides à ce niveau. Sous cette hypothèse, deux règles peuvent être comparables même si leurs mesures sont différentes. Ainsi, nous avons défini des relations d'ordre partiel dans chacune des bases des règles M_{GK}-valides à l'exception de la base des règles négatives exactes et, à partir de ces relations, nous avons pu définir trois semi-bases des règles valides selon la mesure M_{GK}. Dans la nouvelle base des règles négatives exactes, nous n'avons pas pu créer une relation d'ordre qui soit stable dans l'ensemble des règles dérivées à cause des propriétés des axiomes d'inférence utilisés pour la dérivation des autres règles.
Il est à noter que par rapport à une base Support-M_{GK} valide, la considération de la semi-base correspondante peut réduire considérablement le nombre des règles à interpréter et regroupe uniquement les règles les plus informatives, relativement au niveau de confiance préalablement fixé par l'utilisateur. Ce type d'ensemble des règles est très utile en pratique, du moins au début d'une analyse, même si ce n'est pas tout à fait une base, puisque qu'avec cet ensemble, certaines règles valides ne pourraient pas être retrouvées ; car ainsi, on a une vue rapide de l'information majeure contenue dans le contexte étudié. Ensuite, pour l'approfondissement et pour être plus exhaustif, il est toujours possible de revenir aux nouvelles bases des règles de départ.

7.4 Algorithmes d'extraction des nouvelles bases

Nous avons proposé des algorithmes d'extraction des nouvelles bases des règles M_{GK}-valides. Ces algorithmes prennent en entrée des motifs particulier comme les fermés, les générateurs et les bordures positives. On sait que des algorithmes d'extraction de ces motifs particuliers sont disponibles dans la littérature. Nous avons jugé pertinent de les utiliser et considérer ces motifs particuliers à l'entrée de nos algorithmes. Par rapport aux anciens algorithmes d'extraction des bases des règles, nous avons apporté deux améliorations. D'abord, pour le choix de la valeur de référence pour valider ou non une règle ; nous avons inclus dans nos algorithmes le calcul de valeur critique de la mesure M_{GK}. Autrement dit, pour un niveau de confiance α, fixé par l'utilisateur, la validation d'une règle r passe par la comparaison de la valeur de $\mathrm{M}_{GK}(r)$ à la valeur critique $M_{GK}^{1-\alpha}$, mais non pas à la valeur α comme c'était le cas auparavant. Ensuite, le choix des motifs à l'entrée constitue une grande différence entre les anciens et les nouveaux algorithmes. La plupart des algorithmes d'extraction des bases des règles sont conçus à partir des motifs fréquents. Cette pratique n'est plus applicable dans l'extraction des bases des règles négatives puisqu'on extrait les règles négatives à partir des motifs positifs et pour un motif fréquent X, son motif négatif associé $\overline{X}$ n'est pas forcément fréquent et inversement, pour un motif non fréquent X, son motif négatif associé $\overline{X}$ peut être fréquent. Nous avons donc utilisé des motifs de support non nul à l'entrée des algorithmes d'extraction des règles négatives. Ainsi, nous avons pu proposer des algorithmes d'extraction

des bases et semi-bases des règles valide selon la mesure M_{GK}. Par rapport aux problématiques de départ et aux résultats de nos études, plusieurs pistes des travaux peuvent être dégagées, nous allons en citer quelques unes.

7.5 Conception d'un outil

Étant donné que beaucoup de chercheurs en didactique des mathématiques ne sont pas des spécialistes en analyse des données, telle qu'elle est présentée dans ce rapport, il est difficile, voire impossible pour un non-spécialiste d'exploiter cet outil d'analyse des données. Une suite logique de notre travail est donc de concevoir un outil facile à utiliser et qui permet d'effectuer une extraction des règles d'association valides selon la mesure M_{GK}.

7.6 Introduire un outil de recherche de causalité dans un Environnement Informatique d'Apprentissage Humain (EIAH)

Supposons qu'un élève échoue face aux activités de recherche de l'ensemble de définition d'une fonction numérique d'une variable réelle et qu'on veuille lui apporter de l'aide. D'abord, il faudra déterminer pourquoi il a échoué. La réponse à cette question dépend de l'état de connaissance de l'élève. Peut-être que l'élève a un problème sur la manipulation des expressions littérales (comme le développement par exemple), ça peut être un problème de manipulation des signes ou encore un problème sur la résolution d'équation ou d'inéquation, etc. C'est à dire que le fait de ne pas atteindre un objectif spécifique peut être la conséquence de plusieurs problèmes différents. Autrement dit, l'unique fait de savoir qu'un élève n'a pas atteint un objectif dans une activité n'est pas suffisant pour déterminer une aide adéquate à l'apprenant. Un enseignant en présentiel peut comprendre la difficulté « de chacun de ses élèves » en analysant les historiques d'apprentissages des apprenants et agir en conséquence. Cette capacité d'analyse et d'adaptation constitue un point fort de l'être humain. Il est donc souhaitable qu'un Environnement Informatique d'Apprentissage Humain (didacticiel, MOOC...) puisse être doté de cette capacité d'adaptation ou d'individualisation. Selon (Lafarge, 2007) « ...nous devons doter notre système d'une certaine capacité à analyser les erreurs en termes de connaissances en jeu, de processus et de stratégies ». À partir des données sur les apprenants, il est possible d'établir des liens implicatifs entre les connaissances, les erreurs, les situations (Croset, 2009). Les techniques de fouille des données sont de plus en plus utilisées pour l'analyse des interactions des apprenants (Merceron et Yacef, 2008). C'est ainsi qu'il nous semble intéressant de réfléchir sur l'élaboration des techniques permettant d'inclure et d'exploiter les concepts sur les bases des règles M_{GK}-valides dans la conception d'un EIAH.

7.7 Qualité des mesures et des règles

À l'issue de l'analyse des données, une règle valide conduit à une prise de décisions et une décision ne doit pas être trop sensible aux fluctuations des données. En 2010, Le Bras *et al.* ont conçu une définition formelle de la notion de robustesse des règles. Par rapport à ces idées, il peut être intéressant d'analyser la sensibilité d'une règle valide selon la mesure M_{GK} face à des fluctuations des données. Cette étude pourrait apporter des informations supplémentaires sur la stabilité des décisions prises suite à l'extraction des règles valides.
Enfin, les études que nous avons présentées dans ce rapport concernent les données binaires c'est-à-dire lorsque chaque variable n'a que deux valeurs possibles : 1 ou 0 (présent ou absent chez chacun des individu ou transaction). Pourtant, bien des situations nécessitent la considération de variables non binaires. C'est ainsi qu'on peut se poser les mêmes questions que dans la situation binaire, on peut citer par exemple la conception des bases des règles dans le contexte quantitatif.

Références bibliographiques

Agrawal, R., Imielinski, T. et Swami, A. (1993). Mining association rules between sets of items in large databases. *In* Buneman, P. et Jajodia, S., éditeurs : *ACM SIGMOD International Conferance on Management of Data*, volume 22, pages 207 – 216, Washington, USA. 4

Artigue, M. (2008). L'influence des logiciel sur l'enseignement des mathématiques : contenus et pratiques. *In Actes du séminaire national : Utilisation des outils logiciels dans l'enseignement des mathématiques*. 1

Artigue, M. (2011). *Les défis de l'enseignement des mathématiques dans l'éducation de base*. UNESCO. 1

Bastide, Y., Taouil, R., Pasquier, N., Stumme, G. et Lakhal, L. (2000). Mining frequent patterns with counting inference. *ACM SIGKDD Explorations Newsletter*, 2(2): 66–75. 117

Bastide, Y., Taouil, R., Pasquier, N., Stumme, G. et Lakhal, L. (2002). Pascal : un algorithme d'extraction des motifs fréquents. *Techniques et Sciences Informatiquess*, 21(1):65–95. 28, 35, 65

Bemarisika, P., Ramanantsoa, H., Totohasina, A. et Ramifidisoa, L. (2012). Résolution d'équation polynômiale par utilisation des logiciels excel et maxima. *In Colloque international sur les TIC*. 1

Blanchard, J., Kuntz, P., Guillet, F. et Gras, R. (2004). Mesure de la qualité des règles d'association par l'intensité d ?implication entropique. 26

BRAGA, E. D. M. (2009). *Enseignement apprentissage de la statistique, TICE et environnement numérique de travail*. Thèse de doctorat, Université Lumière Lyon 2. 1

Cadot, M. (2006). *Extraire et valider les relations complexes en sciences humaines : statistiques, motifs et règles d'association*. Thèse de doctorat, Université de Franche-Comté. 4

Cariou, J.-Y. (2010). *Former l'esprit scientifique en privilégiant l'initiative des élèves dans une démarche s'appuyant sur l'épistémologie et l'histoire des sciences*. Thèse de doctorat, Université de Genève. 17

Couturier, R. et Almouloud, S. A. (2009). Historique et fonctionnalités de chic. *Revue des Nouvelles Technologies de l'Information*, Analyse Statistique Implicative - Une méthode d'analyse de données pour la recherche de causalités, RNTI-E-16:279–294. 14

Croset, M.-C. (2009). *Modélisation des connaissances des élèves au sein d'un logiciel éducatif d'algèbre. Etude des erreurs stables inter-élèves et intra-élèves en terme de praxis-acte.* Thèse de doctorat, Université de Joseph Fourrier - Grenoble I. 143

Diatta, J., Henri, R. et Totohasina, A. (2007). Towards a unifying probabilistic implicative normalized quality measure for association rules. *In Quality Measures in Data Mining*, pages 237–250. 30

Emprin, F. (2008). Formation initiale et continue pour l'enseignement de mathématiques et les tice : cadre d'analyse des fformations et ingénieries didactiques. *In Actes du seminaire national de didactique de mathématiques, ARDM et IREM de Paris7.* 1

Feno, D. R. (2007). *Mesures de qualité des règles d'association : normalisation et caractérisation des bases.* Thèse de doctorat, Université de la Réunion. 5, 58, 61, 74, 78, 82, 83, 85

Feno, Daniel, R., Diatta, J. et Totohasina, A. (2006). Une base pour les règles d'association d'un contexte binaire valides au sens de la mesure de qualité mgk. *In Comptes Rendus des 13ème Rencontres de la Société Francophone de Classification*, pages 105–109. Université Paul Verlaine-Metz. 58, 60, 61, 65, 70, 71, 83, 85

Gasmi, G., Yahia, S. B., Nguifo, E. M. et Slimani, Y. (2006). IGB : une nouvelle base générique informative des règles d'association. *Revue I3 (Information-Interaction-Intelligence)*, 6(1):31–67. 35, 45, 53, 57, 58, 65, 71, 76

Gras, R. (1996). *L'implicartion statistique.* La pensée sauvage Editions, Grenoble. 4

Gras, R., Kuntz, P. et Briand, H. (2001). Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données. *Mathématiques et sciences humaines. Mathematics and social sciences*, (154). 9, 14

Gras, R. et Larher, A. (1992). L'implication statistique, une nouvelle méthode d'analyse de données. *Mathématiques et sciences humaines*, 120:5–31. 4

Gras, R. et Régnier, J.-C. (2009). Fondements théoriques de l'analyse statistique implicative. *Revue des Nouvelles Technologies de l'Information*, Analyse Statistique Implicative - Une méthode d'analyse de données pour la recherche de causalités, RNTI-E-16:17–130. 4, 14, 21

Grissa, D. (2013). *Etude comportementale des mesures d'intérêt d'extraction de connaissances.* Thèse de doctorat, Université Blaise Pascal-Clermont-Ferrand II ; Université de Tunis-El Manar (Tunisie). 26, 30

Guillaume, S., Grissa, D. et Nguifo, E. M. (2010). Propriétés des mesures d'intérêt pour l'extraction des règles. *Qdc2010, qualité des données et des connaissances.* 30, 31, 32

Hafida, A. (2013). Expansion de la représentation succincte des générateurs minimaux. Maîtrise en informatique de gestion, Université du Québec à Montréal. 117

Hamrouni, T., Yahia, S. B. et Nguifo, E. M. (2011). Construction efficace du treillis des motifs fermés fréquents et extraction simultanée des bases génériques de règles. *Mathématiques et sciences humaines. Mathematics and social sciences*, (195):5–54. 35, 45

Hamrouni, T., Yahia, S. B. et Slimani, Y. (2005). Prince : Extraction optimisée des bases génériques de régles sans calcul de fermetures. *In 23rd French Conference Informatique des Organisations et Systémes d'Information et de Décision (INFORSID'05)*, pages 353–368, grenoble, France. Presse Universitaire de Grenoble. 35, 76, 117

Houston, K. (2009). *How to think like a Mathematician*. Cambridge University press. 20

Karsenti, T. (2009). *Intégration pédagogique des TIC : Stratégies d'action et pistes de réflexions*. Ottawa, IDRC. 2

Karsenti, T., Collin, S. et Harper-Merrett, T. (2012). *Intégration pédagogique des TIC : Succès et défis de 100+ écoles africaines*. Ottawa, IDRC. 3

Khvilon, E., Patru, M., Anderson, J. et Weert, T. V. (2004). *Technologies de l'Information et de la Communication en Éducation : Un programme d'enseignement et un cadre pour la formation continue des enseignants*. UNESCO. 14

Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H. et Verkamo, A. I. (1994). Finding interessing rules from large sets of discovered association rules. *In 3rd International Conference on Information and Knowledge (CIKM'94)*, pages 401–407. 35

Lafarge, V. (2007). L'analyse des erreurs comme outil dans la détermination des aides proposées par un EIAH. *Alsic En ligne*, Volume 10(N°1). 143

Lallich, S. et Teytaud, O. (2004). Evaluation et validation des règles d'association. 27

Le Bras, Y., Meyer, P., Lenca, P. et Lallich, S. (2010). Mesure de la robustesse de règles d'association. *proceedings of the QDC*. 26, 27, 144

Le Floc'h, A., Fisette, C., Missaoui, R., Valtchev, P. et Godin, R. (2003). JEN : un algorithme efficace de construction de générateurs pour l'identification des règles d'association. *Numéro spécial de la revue des Nouvelles Technologies de l'Information*, 1(1):135–146. 35, 117

Lenca, P., MEYER, P., Picouet, P., Vaillant, B. et Lallich, S. (2003). Critères d'évaluation des mesures de qualité des règles d'association. *Revue des Nouvelles Technologies de l'information (RNTI)*, RNTI-1:123–134. 26

Luxenburger, M. (1991). Implication partiel das un contexte. *In Mathématiques, Informatique et Sciences Humaines*, volume 29, pages 35–55. 47, 51

Merceron, A. et Yacef, K. (2008). Règles d'association et analyse d'interaction d'apprenants : mesures d'intérêt. *In L'apprenant et ses nouvelles attentes, au coeur des TICE*. 143

Nguifo, S. B. Y. M. (2004). Approches d'extraction de règles d'association basées sur la correspondance de galois. 117

Ottaviani, M.-G. et Zannoni, S. (2001). Implication statistique et recherche en didactique. utilisation d'un outil non symétrique d'analyse de données pour l'interprétation des résultats d'un test d'évaluation. *Mathématiques et sciences humaines*, (154). 24

Pasquier, N. (2000a). *DATAMINING : Algorithmes d'extraction et de réduction des règles d'association dans les bases des données*. Thèse de doctorat, Université Clermont-Ferrand II. 35, 45, 46, 52, 54, 65, 76, 79, 82, 118

Pasquier, N. (2000b). Extraction de bases pour les règles d'association à partir des itemsets fermés fréquents. *In INFORSID'2000 Congress*, pages 56–77, Lyon, France. Prix Jeune Chercheur de l'association INFORSID. 28, 52, 57, 83

Pasquier, N., Bastide, Y., Taouil, R. et Lakhal, L. (1999a). Closed set based discovery of small covers for association rules. *In Proc. 15emes Journees Bases de Donnees Avancées, BDA*, pages 361–381, Bordeaux,France. 49

Pasquier, N., Bastide, Y., Taouil, R. et Lakhal, L. (1999b). Discovering frequent closed itemsets for association rules. *In Proc. ICDT conf., LNCS 1540*, pages 398–416. 49

Pasquier, N., Bastide, Y., Taouil, R. et Lakhal, L. (1999c). Efficient mining of association rules using closed itemset lattices. *Information Systems*, 24:25–46. 49

Pelgrum, W. J. et Law, N. (2004). *Les TIC et l'éducation dans le monde : tendances, enjeux et perspectives*. UNESCO. 2

Ramanantsoa, H., Bemarisika, P., Totohasina, A. et Ramifidisoa, L. (2012). Enseignement de limite d'une fonction et TIC au niveau secondaire. *In Colloque international sur les TIC*, ENS Ampefiloha Antananarivo. 1

Ramanantsoa, H. et Totohasina, A. (2014). Une stratégie d'intégration pédagogique des TIC dans l'enseignement des mathématiques à Madagascar : A strategy for integration of ICT in teaching mathematics in Madagascar. *Frantice.net*, (9). 14

Ramanantsoa, H. et Totohasina, A. (2015). Note sur les bases des règles valides selon la mesure M_{GK}. *In XXII ème Rencontre de la société Francophone de Classification*. 78

Rioult, F., Zanuttini, B. et Crémilleux, B. (2010). Nonredundant generalized rules and their impact in classification. *In Advances in Intelligent Information Systems*, pages 3–25. Springer. 78

Salleb, A. (2003). Recherche de motifs fréquents pour l'extraction de regles d'association et de caractérisation. *These de doctorat, Laboratoire d'Informatique Fondamentale d'Orléans LIFO, Université d'Orléans*. 42, 117

Srikant, R. et Agrawal, R. (1996). Mining sequential patterns : Generalizations and performance improvements. *In Biennial International Conference on Extending Database Technology (EDBT'96)*, numéro 5, pages 3–17, Avignon, France. 19

TOTOHASINA, A. (2008). *Contribution à l'étude des mesures de la qualité des règles d'association : normalisation sous cinq contraintes et cas de M_{GK} : propriétés, bases composites des règles et extension en vue d'applications en statistique et en sciences physiques*. HDR, Université d'Antsiranana. 19, 32, 60, 98

TOTOHASINA, A. et FENO, Daniel, R. (2008). De la qualité des règles d'association : étude comparative des mésures M_{GK} et confiance. *In CARI'2008*, pages 561–568. 71

TOTOHASINA, A. et RALAMBONDRAINY, H. (2005). Ion : a pertinent new measure for mining information from many types of data. *In SITIS*, pages 202–207. 5

TOTOHASINA, A., RALAMBONDRAINY, H. et DIATTA, J. (2004). Une vision unificatrice des mesures de la qualité des règles d'association booléennes et un algorithme efficace d'extraction des règles d'association implicative. 5, 30, 32, 71, 72

VAILLANT, B. (2006). Mesurer la qualité des règles d'association : études formelles et expérimentales. *France, These de PhD, École Nationale Supérieure des Télécommunications de Bretagne et Université de Bretagne sud*. 27

VAILLANT, B., LENCA, P. et LALLICH, S. (2004). Etude expérimentale de mesures de qualité de règles d'associations. *In EGC*, volume 4, pages 341–352. 27

VAILLANT, B., MEYER, P., PRUDHOMME, E., LALLICH, S., LENCA, P. et BIGARET, S. (2005). Mesurer l'intérêt des règles d'association. *In EGC (Ateliers)*, pages 421–426. 27

WU, X., ZHANG, C. et ZHANG, S. (2004). Efficient mining of both positive and negative association rules. *ACM Transactions on information Systems*, 3:381–405. 5

ANNEXES

Annexe A

Extraits des données recueillies

	Out_Tice s	Out_TN s	Mait_Pre_2	Util_Def_F_aut	Df_Calc_For	Df_LecG_For	LecG_ia_For
E_1	0	1	0	0	0	0	1
E_2	0	1	1	1	0	0	1
E_3	0	1	1	1	1	0	0
E_4	0	1	1	1	1	0	1
E_5	0	1	1	0	1	0	1
E_6	0	1	0	0	0	1	1
E_7	0	1	0	0	0	1	1
E_8	0	1	0	0	0	0	0
E_9	0	1	1	0	0	0	1
E_10	0	1	0	0	0	0	0
E_11	0	1	1	0	0	0	0
E_12	0	1	1	1	1	0	1
E_13	0	1	1	0	0	1	1
E_14	0	1	0	0	0	0	1
E_15	0	1	1	0	0	0	1
E_16	0	1	1	1	1	0	1
E_17	0	1	0	0	0	0	0
E_18	0	1	0	0	0	0	0
E_19	0	1	0	0	0	0	1
E_20	0	1	0	0	0	0	1
E_21	0	1	1	0	0	0	0
E_22	0	1	1	1	1	1	1
E_23	0	1	1	1	0	1	0
E_24	0	1	1	0	0	1	1
E_25	0	1	0	0	0	0	1
E_26	0	1	0	0	0	1	1
E_27	0	1	1	1	1	1	1
E_28	0	1	1	1	0	0	1
E_29	0	1	1	0	0	0	1
E_30	0	1	1	0	0	1	0
E_31	0	1	1	0	0	0	0
E_32	0	1	1	1	1	1	1
E_33	0	1	1	1	1	0	0
E_34	0	1	1	1	1	0	1
E_35	0	1	1	0	0	1	1
E_36	0	1	1	0	0	0	1
E_37	0	1	0	0	0	0	1
E_38	0	1	0	0	0	0	0
E_39	0	1	1	0	0	0	0
E_40	0	1	1	0	0	0	1
E_41	1	0	1	1	0	1	1
E_42	1	0	1	0	0	0	0
E_43	1	0	1	1	0	0	0

	LecG_eq_For	LecG_ine_For	LecG_ext_For	Df_Sit_Pro	LecG_ia_Pro	LecG_eq_Pro	LecG_ine_Pro	LecG_ext_Pro
E_1	0	0	1	0	1	0	0	1
E_2	0	0	0	0	1	0	0	1
E_3	0	0	0	0	1	1	1	1
E_4	1	0	1	1	0	0	0	0
E_5	1	1	1	0	1	1	0	1
E_6	1	0	1	0	1	0	0	0
E_7	0	0	1	0	1	1	0	0
E_8	0	0	1	0	1	1	0	1
E_9	1	0	1	0	1	0	1	0
E_10	0	0	1	0	1	0	0	0
E_11	0	0	0	0	0	0	0	0
E_12	0	1	1	1	1	1	0	1
E_13	0	0	0	0	0	0	0	0
E_14	0	0	1	0	1	1	1	1
E_15	0	0	1	0	1	1	0	1
E_16	1	1	1	0	1	0	0	0
E_17	0	0	0	0	1	0	0	1
E_18	0	0	1	0	0	0	0	1
E_19	0	0	0	0	0	0	0	0
E_20	0	0	0	0	1	0	0	1
E_21	0	0	1	0	1	1	0	1
E_22	1	0	1	1	1	1	1	1
E_23	0	0	1	0	1	0	0	0
E_24	1	1	1	0	1	1	1	1
E_25	0	0	1	0	1	0	0	1
E_26	0	0	1	0	1	0	0	0
E_27	0	0	0	0	1	1	1	1
E_28	1	0	1	0	1	1	0	0
E_29	0	0	1	0	1	0	0	1
E_30	0	0	0	0	0	0	0	0
E_31	0	0	0	0	1	1	0	0
E_32	1	1	1	0	1	1	1	1
E_33	0	0	1	0	1	1	0	1
E_34	0	0	1	0	1	1	0	0
E_35	1	1	1	1	1	1	1	1
E_36	0	0	1	0	1	0	0	1
E_37	0	0	1	0	1	1	0	0
E_38	0	0	0	0	1	0	0	1
E_39	0	0	0	0	1	1	0	0
E_40	0	0	0	0	1	1	0	0
E_41	1	0	1	0	1	1	0	1
E_42	1	1	1	1	1	1	1	0
E_43	0	0	0	0	1	1	1	1

Résumé

L'analyse statistique implicative (ASI) est un outil de découverte des liens de cause à effet très adapté aux recherches en didactique de disciplines. Avec l'ASI, l'analyse de nos données ainsi obtenues à l'issue d'une expérimentation nous a montré que l'utilisation simultanée et équilibrée des TIC et des tableaux est l'un des moyens qu'on peut utiliser pour atteindre les objectifs de l'enseignement de mathématiques.
La fiabilité des résultats d'une analyse des données dépend fortement de l'outil mathématique utilisé. Dans le cas de l'ASI, elle dépend de la qualité des mesures utilisées. Nous avons choisi de travailler avec le couple de mesures $Support - \mathrm{M}_{GK}$, la mesure M_{GK} ayant son unique composante active qui est implicative, propriété tant souhaitée en ASI.
Le nombre des règles valides par une quelconque mesure peut être très élevé parce que la plupart sont des règles redondantes. Cela complique les interprétations. La génération d'un ensemble minimal des règles contenant toutes les règles les plus informatives que l'on appelle communément bases des règles est un moyen pour contourner ce problème. Avec le couple de mesures $Support - \mathrm{M}_{GK}$, nous avons défini des bases plus pertinentes et la notion des sémi-bases des règles. Ensuite, nous avons proposé des algorithmes d'extraction des bases et des semi-bases.

Mots-clés : Règles d'association, Mesure de qualité, Bases et sémi-base des règles, TICE.

Abstract

The Statistical Implicative Analysis (SIA) is a very suitable tool for discovering causes and effects in discipline education research . With SIA, the analysis of our data such obtained after an experiment has shown that the simultaneous and balanced use of ICT and blackboard or whiteboard is a fructufull way we can use to achieve the teaching mathematics objectives.
The reliability of results of data analysis strongly depends on the used mathematical tool. In case of the SIA, it depends on the quality of the used measurements. We choose dealing with both measurement Support and M_{GK}, because the unique active component of M_{GK} is implicative, such property being requered in SIA. The number of valid rules by any measure can be very huge because of many redundant rules. This situation complicates interpretation. Generating a minimum set of rules containing all the most informative rules that are commonly called bases of the rules is a benefit way to around this drawback. With couple of measurement Support-M_{GK}, we defined the new concept of bases and semi-bases of rules. Then we proposed bases and semi-bases extraction algorithms.

Keys-words : Association rules - Quality of measures - Base - ICT in Education