Contribution à l'étude des mesures de la qualité des règles d'association : normalisation sous cinq contraintes et cas de MGK : propriétés, bases composites des règles et extension en vue d'applications en statistique et en sciences physiques.

André Totohasina

Habilitation à Diriger des Recherches

Spécialité

Mathématiques et Informatique

Université d'Antsiranana
Madagasikara

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.

présentée par

Dr TOTOHASINA André

Maître de Conférences

soutenue le 21 Février 2008 devant le jury composé de

M. Fabien CAMPILLO, Chargé de Recherche Habileté (INRIA, France), Rapporteur
M. Victor HARISON, Professeur (Univ. d'Antanànarivo, Madagasikara) Rapporteur
M. Andriantiana Bertin Olivier RAMAMONJISOA, Professeur (Univ-Fianarantsoa, M/kara), Rapporteur
M. Edouard ALIDINA, Professeur (Université d'Antsiranana, Madagasikara), Examinateur
M.Michel DAGUENET, Professeur Émérite (Univ. de Perpignan, France), Examinateur
M. Jean DIATTA, Professeur(Université de La Réunion, France), Examinateur
M. Germain Eugène RANDRIAMBELOSOA, Professeur (Univ. d'Antanànarivo, M/gasikara) Examinateur
M. Henri RALAMBONDRAINY, Professeur (Université de La Réunion, France), Directeur

Dédicaces

À la mémoire de mon père qui a su me donner le goût des études. Je me souviens toujours de ces devoirs de calculs et de français qu'il me confiait de traiter à la maison à Antsahanibakôko, en brousse à douze kilomètres du chef lieu de district de Mandritsara, alors qu'il devait rejoindre sa garderie dans le village de Maroandriagna situé à deux kms de mon village. Quels encouragements de sa part en cas de traitement effectif à son retour! Jouer m'était toujours permis. Mais des coups de sa ceinture aux jambes, parfois des tirs d'oreilles m'attendaient sinon!...

Ainsi, un peu tard seulement, il m'inscrivait à l'ecole catholique d'Antanjagno, encore une garderie en fait, située à un km de mon village, mais de l'autre côté de la grande rivière Mangarahara : j'y étais directement inscrit en classe de 10ième. L'année d'après, je participais au concours d'entrée en 8ième de l'école de la mission catholique de la ville de Mandritsara. Mes parents m'y confiaient à une tante et se contentaient de m'apporter provisions en vivres de temps en temps. Les jours, les semaines, les mois passaient. J'étais parmi les lauréats du CEPE de la ville en fin du cours moyen deuxième année, malgré maintes tentatives de renvoi faute de régularité sur le paiement d'écolage mensuel! Ma vie scolaire continuait ainsi loin des parents jusqu'à l'obtention du BEPC. L'admission en seconde dans l'unique Lycée public, le Lycée Victor Miadana , se faisait sur concours pour les sept districts de la région Sofia. Je le réussisais et je faisais même partie des 10 premiers! Ce qui m'encourageait à participer au concours national, contingenté à 10 garçons et 8 filles par province, d'entrée en "seconde normale" de l'école normale des instituteurs de Mahamasina Antanànarivo, la capitale de Madagascar : j'y étais admis dans la section spéciale de "seconde AC". Depuis 1973, je suis devenu un normalien et le continue jusqu'à l'enseignement supérieur (Institut national supérieur de recherche et de formation pédagogique, ENS de Fianarantsoa, recherche en didactique de maths, permanent à l'ENS d'Antsiranana avec 7 ans de direction), aussi bien en tant qu'étudiant qu'en qualité d'enseignant chercheur.

À la mémoire de ma mère bien évidemment! Elle était illéttrée, mais savait bien me gérer avec affection. Je regrette beaucoup sa disparition subite, suite à un empoisonnement dans son café, alors qu'elle ramenait les restes de sa mère dans son village natal et maternel Anjanabôrogno, Bealanana. J'apprenais ce douloureux et triste événement à Rennes, pendant mon séjour de préparation de thèse de doctorat, fin septembre 1990. Je passais toutes mes vacances scolaires chez ses parents en brousse à Ambalantsôtry, situé à 20 km de Mandritsara.

À la mémoire de ma belle-grand-mère RASOAVELO qui m'avait bien reçu à Bealanana.
À la mémoire de ma grand-mère maternelle VOLATIANA, originaire d'Anjanabrogno Bealanana, qui m'accordait beaucoup d'affection à Ambalantsôtry, durant mes grandes vacances scolaires. Elle m'offrait un jeune taureau pour financer mes premiers équipements avant de rejoindre l'école normale de Mahamasina en 1973.

À mes filles Nathalie, Ninah, Lucie et à mon garçon Gérald Stone.
À ma Chère bien aimée Épouse RAZANAMISY Rasoambololona.

Ne peut-on pas parler de l'éducation mathématique?

Les mathématiques aiment et recherchent les choses simples, c'est-à-dire la simplicité. Pour s'en convaincre, il suffit de retracer les fréquentes questions de simplification posées par nos institeurs du primaire et par nos profeseurs du collège et du lycée, du genre :

1. **En primaire :** Simplifier les fractions $\frac{49}{14}$, $\frac{36}{840}$, ..., etc.
2. **En secondaire :** Simplifier les fractions rationnelles $\frac{X^2-3x+2}{X^2-1}$, $\frac{5X^2-3x-2}{X^3-1}$, ..., etc.

Par ailleurs, on nous a toujours dit qu'en mathématiques, parmi plusieurs démonstrations possibles, ou parmi plusieurs solutions à un problème, la plus concise, c-à-d la plus simple, est la meilleure!

Cette philosophie d'aspiration aux choses simples demeure en informatique, qui est d'ailleurs une discipline parente des mathématiques! On y recherche l'algorithme de plus bas degré de complexité, c-à-d le moins complexe possible!

Ceci n'est pas sans inconvénient, quant au comportement d'un mathématicien comparativement à celui d'un collègue des sciences humaines, cette tendance à simplifier les choses gêne souvent nos administrateurs qui sortent d'un autre moule!

Cependant, heureusement, à force d'être entraîné par les multiples changements de terminologies à travers les divers théories et cadres mathématiques (algèbre, analyse, probabilités, statistiques, géométrie, etc.) tout au long de la formation, un mathématicien incarne la flexibilité (par la capacité d'adaptation au milieu ambiant et ce avec une objectivité bien scientifique) tant recherchée dans la vie professionnelle et sociale!

Toutefois, cela ne signifie pas que les mathématiques et l'informatique sont simples!

Mais, les nuits portent conseils!
Les échanges entre paires sont enrichissants!
La littérature spécialisée informe et réveille!
En classe ou dans un bureau de chercheur, le tableau est magique!
... Et souvent, c'est ainsi qu'une bonne idée sur la résolution d'un problème se fait découvrir.

Remerciements

Mes premiers remerciements vont à Henri RALAMBONDRAINY, Professeur à l'Université de La Réunion, qui m'a toujours réservé un accueil chaleureux lors de mes séjours scientifiques au sein de son équipe ECD du laboratoire IREMIA, depuis 1997, puis chaque année depuis 2001 pendant les 2 ou 3 mois de séjours.
Mes remerciements vont également à Jean DIATTA, Professeur à l'Université de La Réunion, avec qui j'ai co-encadré en alternances les travaux de thèse du doctorant Daniel Rajaonasy FENO. Je me réjouis et me sens intellectuellement très honoré que ce dernier a bien pu mener ses travaux à termes, il a pu soutenir son doctorat de l'Université de La Réunion à temps, le 1er décembre 2007. Des échanges fructueux ont eu lieu entre lui et moi. Je pense avoir pu réaliser une de ses perspectives de recherche évoquées dans sa thèse d'HDR, à savoir : "Envisager l'intégration d'autres critères de qualité de règles d'association permettant de construire des bases qui contiendraient des règles plus cohérentes par rapport aux bases de Guigues-Duquenne-Luxenburger".

Je tiens également à adresser mes vifs remerciements ainsi que ma profonde gratitude à toutes les personnalités qui composent mon jury d'HDR, notamment aux Professeurs qui ont bien voulu rapporter le présent document de synthèse.

Mes remerciements vont aussi à l'endroit des organismes, à savoir AUF, SCAC de la coopération française, Université de La Réunion à travers les attributions de poste de MCF invité, SARIMA-CIMPA, qui m'ont accordé financement pour mes différentes activités scientifiques qui devaient s'effectuer autour de ma préparation d'HDR : séjours à l'Université de La Réunion, déplacements pour participation à des colloques internationaux.

Je remercie naturellement mes supérieurs hiérarchiques de mon Université d'attache, Madame MANOROHANTA Cécile Marie-Ange, Présidente de l'Université d'Antsiranana, M. ANDRIANIRINA Charles Bernard, Le Vice-Président, Messieurs RABE Tsirobaka et JEANNOT directeurs de l'ENSET, qui m'ont toujours accordé l'autorisation d'absence et de sortie du territoire chaque fois que j'obtenais un financement, soit pour séjourner dans un laboratoire étranger, soit pour participer à des colloques internationaux. Qu'ils trouvent ici toutes mes reconnaissances.

Enfin, et non le moindre, mes vifs remerciements s'adressent à Monsieur Le Professeur Edouard ALIDINA, le Responsable de la Formation Doctorale, qui a bien voulu faciliter la complétion des formalités administratives exigées pour une HDR, ainsi qu'à Monsieur RIZIKY Gen Hiviel Tsiresena, le Doyen de la faculté des Sciences, qui a daigné prendre en charge les indemnités de membres de mon Jury et une bonne partie des frais de déplacements et séjour de certains de mes examinateurs.

Table des matières

Chapitre 1

Thèmes de recherche et Contributions

1.1 Préambule

Après mes trois années de doctorat à l'Université de Rennes I, France, j'ai effectué une année de stage post-doctorale(l'année universitaire 1993/1994) à l'Université du Québec à Montréal (UQAM), Canada, alors que le processus administratif de mon recrutement à l'Université Nord Madagascar, devenue l'actuelle Universié d'Antsiranana, poursuivait son cours.

À l'UQAM, j'étais à cheval entre le Centre Interdisciplinaire de Recherche Appliquée au Développement de l'Enfant (CIRADE), qui abrite des chercheurs en éducation dont les chercheurs en didactique des disciplines, et le département de mathématiques à travers son équipe de Statistique. Je fonctionnais en qualité de personne ressource en traitements et analyse des données (pour faire valoir, entre autres, mes compétences en Analyse Statistique Implicative(ASI), objet principal de ma thèse de doctorat) dans le CIRADE ; parallèlement au CIRADE, je poursuivais ma recherche en ASI au sein de l'équipe statistique du département des mathématiques et suivais des cours de niveau II en statistiques et en probabilités. Deux articles portant sur la méthodologie d'ASI agrémentée de son application à des données de recherche en didactique des mathématiques sont redigées et parues dans deux revues à audience internationale ([GT95b] ; [GT95a]) et un preprint [Tot94] sur les valeurs critiques de l'indice d'implication de R. Gras y ont été élaborés. Cette approche de l'analyse statistique implicative a beaucoup évolué depuis ces dernières années tant sur le plan théorique, avec ses diverses versions et extensions , que pratique ((cf. [GKCG01], [BGBG05], [HGB05a], [HGB05b], [GK07], etc.).

Ainsi, en fin juillet 1994 je réjoignis mon poste à la Faculté des Sciences de l'Université Nord d'Antsiranana, dans le département de Mathématiques et Physique. Ensuite, depuis février 1995 jusqu'en fin septembre 2002, j'ai occupé le poste de directeur de l'École Normale Supérieure pour l'Enseignement Technique(ENSET), établissement de la formation initiale des Enseignants de lycée général ou technique sur les trois filières de Génie Mathémati- ques et Informatique, de Génie Mécanique et de Génie Électrique, jusqu'en fin septembre 2002. Mes activités pédagogiques se répartissent dans les trois établissements de l'université d'Antsiranana comme suit :

à l'ENSET : Analyse 2 et Probabilités & Statistiques (1er cycle, de 1995 à

2002), Algèbre multilinéaire et tenseurs - Programmation linéaire et optimisations - Compléments de probabilités (second cycle de 1995 à ce jour), Anneaux et extensions de corps (second cycle de 1995 à 2003), Analyses statistiques multivariées - Transformations intégrales et EDPs- Introduction aux processus stochastiques (jusqu'en 2003)- Didactique et pédagogie des mathématiques (de 1995 à ce jour) - suivi des stagiaires de cinquième année. Encadrements d'une trentaine d'étudiants sortants pour le CAPEN depuis 1995 à ce jour.

à la Faculté de Sciences : Analyse I - analyse II - probabilités (1994 à 1996) ; Maths IV (2006/2007) ; Modélisations mathématiques (DEA de physique depuis 2005) ; encadrement en cours de deux étudiants de DEA de Mécanique de Fluides et Système Énergétique (PASCAL Petera : Approche statistique d'un problème d'écoulement plan ; puis AMBEONDAHY : Résolution analytique par transformation de Laplace des équations de chaleur à une et à deux dimensions).

à l'École Supérieure Polytechnique : Compléments de mathématiques (transformations intégrales et optimisations) au DEA d'électromécanique depuis 2005.

Avec le concours de deux Professeurs permanents à l'université d'Antanànarivo (Pr Rajoelina Michel et Pr Razafy Andriamampianina (feu) et la collaboration du Professeur Richter Herbert, j'ai monté un DEA de mathématiques (Analyse, Statistiques mathématiques et applications) en 1996-97. Ce qui m'a permis d'assurer un encadrement de DEA soutenu :
Daniel R. FENO : *Position de l'implication statistique de Gras en tant que concept de dépendance statistique par rapport à la classification de Lehmann et application en didactique des mathématiques, D.E.A., Faculté de Sciences de l'Université d'Antanànarivo et Université d'Antsiranana, Madagascar, avril 1999.*

Enfin, avec Pr Jean Diatta, de l'Université de La Réunion, j'ai co-encadré un doctorant qui a également soutenu sa thèse de doctorat[Fen07] le premier décembre 2007 à l'Université de La Réunion.

Depuis 2001, j'ai pu m'associer avec l'équipe ECD du laboratoire IREMIA de La Réunion pour venir y poursuivre en alternances mes activités de recherche, objets de ce mémoire, avec le soutien financier de l'Agence Universitaire de la Francophonie (AUF)(séjour de 2 semaines de prospection) et par celui du Service de Coopération et d'Action Culturelle (SCAC) de la France à deux reprises (2 fois trois mois), grâce aussi à l'obtention d'un poste de Maître de Conférence invité à l'Université de La Réunion(3 mois en 2005 ; 2 en 2006).

Mes travaux de recherche relèvent du domaine de la Fouille de Données (F.D.)(ou " Data mining"). Ils s'inscrivent ainsi dans le cadre de l'Extraction des Connaissances à partir des Bases de Données (E.C.B.D. ou plus simplement E.C.D.) : mesure de la qualité des règles d'association, classification, statistique inférentielle, analyse formelle des concepts et Applications [Vou02]. En matière d'application, mes travaux concernent essentiellement des données de recherche en didactique des mathématiques. Aussi, naturellement (car il n' y a que les professeurs de mathématiques ou les mathématiciens qui soient les mieux placés pour s'occuper du problème de l'apprentissage des mathématiques...!), professant les mathématiques et la didactique des mathématiques, mes travaux touchent la didactique des mathématiques, voire le suivi des étudiants en stage de responsabilité dans un lycée. Ce dernier sujet relève déjà du domaine de la Science de l'Éducation, certes, mais cette flexibilité

envers mon milieu professionnel ne fait que témoigner de l'avantage de l'éducation mathématique résultant systématiquement d'une formation mathématique et d'une formation psychopédagogique de haut niveau pour la formation de l'Homme.

Ce document présente les approches que j'ai développées dans chacun de ces domaines, ainsi que les travaux que j'ai menés après ma thèse de doctorat.

Dans le premier chapitre, je présente le cadre contenant le domaine de mes travaux, mes thèmes de recherche ainsi que mes principales contributions. Dans le second chapitre je fais un état de l'art sur la fouille des règles d'association dans un contexte binaire et j'y définis la notion de mesure probabiliste de la qualité des règles. Dans le troisième chapitre, je développe mon approche sur le concept de mesure de qualité normalisée : des propositions ainsi que le processus de normalisation d'une mesure de qualité y sont donnés. Une étude particulière de la mesure de qualité de Guillaume-Kenchaff [Gui00] notée M_{GK}, y est également présentée : ses proprétés mathématiques et son application pour construire une base composite des règles d'association. M_{GK} va jouer un rôle centrale au sein d'un groupe majoritaire des mesures probabilistes d'intérêt de règle ainsi qu'une possibilité de son extension. Dans le quatrième chapitre, je dégage un certain nombre d'applications de M_{GK} sur le traitement de données quantitatives en vue d'extraire des règles d'association quantitative de diverse types, et sur l'analyse des données complexes. Enfin, une conclusion contenant quelques perspectives de recherche s'inscrivant dans la continuation des travaux présentés dans ce document de synthèse termine ce mémoire.

1.2 Introduction

1.2.1 Préliminaires

S'inscrivant dans le domaine de fouille de données, s'avérant ainsi connexes à mes travaux de thèse de doctorat [Tot92], mes activités de recherche gravitent autour de l'Analyse Statistique Implicative (ASI). Rappelons tout d'abord que la situation fondamentale et fondatrice de l'ASI est la suivante [LGR81]. Un ensemble d'objets ou d'individus est observé par rapport à un nombre fini de variables ou caratères ou items et l'on s'y pose la question de type : "*Dans quelle mesure peut-on considérer qu'instancier la variable ou un ensemble de variables A implique instancier la variable ou un ensemble de variables B ? Autrement dit, les objets ont-ils tendance à être B si l'on sait qu'ils sont A ?* ". Dans les situations naturelles, humaines ou sciences de la vie, les théorèmes (si A, alors B) au sens déductif du terme ne peuvent être établis du fait des exceptions qui les entachent, il est important pour le chercheur et le praticien de "fouiller dans ses données" afin de dégager cependant des règles suffisamment fiables pour pouvoir conjecturer une possible relation causale, une genèse pour décrire, structurer une population d'objets et émettre l'hypothèse d'une stabilité à des fins descriptives et, si possibles prédictives. Mais cette fouille nécessite la mise au point de méthodes pour la guider et pour la dégager du tâtonnement et de l'empirisme. Pour la suite, un ensemble de variables est désigné par un motif.

Depuis un peu plus d'une dizaine d'années, selon la méthode empruntée, cette problématique de recherche des règles d'association valides au sens d'un critère (ou d'une mesure de qualité ou d'intérêt) précis occupe la plupart des travaux sur les techniques de fouille de données.

L'extraction des règles d'association est devenue aujourd'hui l'une des tâches

la plus populaire de la fouille de données, et ce, depuis les travaux de Agrawal et al. [AIS93, AS94]. L'analyse de panier de ménagère est l'une des applications typiques de l'extraction des règles d'association. Elle a pour but de dégager les relations intelligibles entre les attributs dans une base de données. Dans le cas de l'analyse du panier de ménagère, l'extraction des règles d'association permet d'analyser les tickets de caisse des clients particuliers afin de comprendre leurs habitudes de consommation, agencer les rayons du magasin, organiser les promotions, gérer les stocks etc. dans le naturel but d'améliorer le profit. Dans la base de données de vente, une *transaction* consiste à un ensemble d'articles achetés par un client particulier, appelés *items* ou *attributs*. Ainsi, une base de données est un ensemble de transactions qu'on appelle aussi base *transactionnelle*. Dans un tel contexte, une règle d'association est une implication conditionnelle entre des ensembles d'attributs dans une base transactionnelle. En d'autres termes, étant donné un ensemble d'attributs, le but de l'extraction des règles d'association est de découvrir "si l'occurrence de cet ensemble dans une transaction est associée à l'occurrence d'un autre ensemble d'attributs". Par exemple, "80% des clients achetant du lait et du thé achètent aussi du pain, sachant que 60% des clients achètent les trois articles lait, thé et pain" est une règle d'association associant les attributs lait et thé à l'attribut pain. Ce problème est loin d'être trivial, vu le nombre exponentiel du nombre d'attributs de la base transactionnelle. Par ailleurs, très rapidement aujourd'hui, dans les grandes aglomérations, une base transactionnelle comprend des millions de transactions sur des milliers d'attributs.

1.2.2 Le cadre du domaine des travaux

On assiste actuellement à l'existence d'énormes bases des données, grâce au développement des techniques et des capacités de stockage. Pratiquement, tous les domaines de l'activié humaine sont concernés par cette explosion de la collecte des données : commerce pour des besoins de marketing et de gestion de stocks ; industrie pour la gestion de production, la maintenance, la prévention des risques, le contrôle de qualité ; le monde judiciaire pour la détection des fraudes ; médecine ; pharmacologie ; génome ; chimie ; internet pour la navigation et la recherche de l'information, etc. Gérer de tels grands volumes de base de données pose le problème de l'exploitation de l'information potentielle qui y est contenue. Aussi, l'analyse, l'organisation et l'interprêtation des données sont devenues cruciales dans une société moderne.

Ces dernières années, on dispose de l'informatique décisionnelle pour pallier le manque de l'informatique de production. Des entrepôts de données sont construits à cette fin dès le début des années 90 (cf. Gilleron & Tamasi, 2000 ; Benitez-Guerrero et al., 2001). Le traitement en ligne des données (cf. OLAP : On-Line Analytical processing) sont rendues possibles grâce à l'architechture de ces nouveaux moyens de stockage des données. On évolue ainsi du langage traditionnel de requête des bases de données vers de meilleures techniques de production d'analyses assez fines et multidimensionnelles des données. Cela repose sur le principe du cube de données. On construit un modèle multidimensionnel dans lequel les données sont décrites selon les dimensions de l'analyse. Par exemple, les données sur les ventes peuvent être observées sous les trois dimensions comme clients, produits et dates d'achat ; ce qui correspond à de vrai cube effectivement. Naturellement, cette notion de cube de données s'étend à de dimensions supérieures à trois. Toutefois, même ce procédé souffre encore des limites. D'où la naissance du nouveau domaine de recherche

qu'est l'extraction des connaissances à partir des bases de données (E.C.B.D.) ou tout simplement l'extraction des connaissances à partir de données (E.C.D.) qui s'est très vite développé.

D'une utilisation traditionnelle pour des besoins tactiques(gestion de la production), les données sont désormais prises en compte dans les entreprises à des fins stratégiques pour augmenter les ventes, réduire les coûts, mais aussi gérer les risques et prévenir les fraudes. De nombreuses sociétés de services se sont spécialisées dans le conseil dans ces domaines. Outre l'intérêt grandissant du monde économique, l'extraction des connaissances à partir de bases de données trouve également de très nombreuses applications dans les domaines social, scientifique et industriel. Le développement de techniques efficaces croît parallèlement à l'apparition de nouveaux besoins, comme, entre autres, la prise en compte de nouveaux types de données.

Bref, l'E.C.D. a deux buts différents : la description et la prédiction. L'objectif descriptif vise à donner une vue globale des données en les regroupant dans des ensembles le plus possible homogènes ; ainsi, les données deviennent plus faciles à appréhender. L'objectif prédictif consiste à estimer les informations a priori inconnues. En général, l'ECD vise à extraire des informations potentiellement utiles à partir de grands volumes de données. Aussi est-il légitime que les connaissances mises en évidence pourrait venir enrichir la base de connaissances d'un système de base de connaissances. Par conséquent, une partie de l'acquisition des connaissances pourrat être automatisée ou plutôt semi-automatisée, car, en fait sur le plan pratique le processus d'extraction de connaissances est avant tout un procédé itératif et intéractif.

Signalons que l'expression "extraction de connaissances à partir de bases de données" vient de la traduction de l'expression anglaise "knowledge discovery in databases (KDD)". Ce terme a été introduit par Piatetsky et Shapiro lors d'un des workshops de la conférence IJCAI'89 [PS91a], [FPSM92]. L'ECD se trouve au confluent de nombreuses autres domaines de recherche : Statistiques [EP96], bases de données, entrepôt de données, interfaçage homme-machine et visualisation, reconnaissance des formes, représentation des connaissances, apprentissage automatique, réseaux de neurones et analyse des données. Elle emprunte à ces différentes disciplines les outils théoriques ainsi que ceux pratiques dont elle a besoin. L'ECD travaille à partir de données. Il s'avère alors naturel que l'étude des bases ainsi que des entrepôts de données constitue une base. Pour analyser et représenter les données, l'ECD utilise des méthodes développées dans les domaines de statistiques et de l'analyse des données, mais aussi de l'intelligence artificielle. L'ECD s'intéresse à la présentation des résultats tout au long du processus et donc aux techniques de visualisation et à l'interfaçage homme-machine.

1.2.2.1 Les étapes du processus d'ECD

L'ECD comprend tout un cycle de découvertes d'informations. Ce processus est complexe et se déroule en plusieurs étapes. La littérature atteste diverses variantes de nombre d'étapes constituant le processus d'ECD : 5, 6, 7, ou 8 [KM06]. Nous retenons l'ECD à 6 étapes à savoir (cf. Fayyad-Piatetsky-Smyth96 1996) [FPSS96] : la compréhension du problème ou de la problématique (car tout processus d'extraction de connaissances répond à un besoin, tout comme la construction d'un système de base de connaissances), la sélection des données à partir d'une base de données,

le prétraitement des données cibles, la transformation des données prétraitées, la fouille de données transformées, l'interprétation /évaluation des modèles issus de la fouille de données en termes de connaissances. Ces différentes phases de l'ECD sont sommairement expliquées dans les lignes qui suivent.

ECD-1 : **Compréhension du problème :** La compréhension de la problématique permet de cibler les buts et guide ainsi la suite du processus. Pour ce faire, on a souvent besoin de bien discuter intensivement avec l'expert utilisateur des données et des connaissances extraites [LHCM00], [Bri04].

ECD-2 : **Sélection des données :** Disposant de très grandes bases de données, il n'est pas nécessaire d'examiner l'ensemble de toutes les données, car l'intégralité de données ne correspond pas forcément au type des données intéressant par rapport aux buts fixés. La sélection de données permet d'élaborer un jeu d'essai adapté au cadre de l'étude. Ainsi, dans le domaine des statistiques, on forme un échantillon repré- sentatif de la population étudiée. Cette étape de sélection peut être précédée d'une étape de recensement des données en cas de non existence au préalable.

ECD-3 : **Pré-traitement :**
Afin d'essayer d'améliorer leur qualité en termes de consistance, complé- tude, précision, homogenéité, etc., on doit pré-traiter le sous-ensemble représentatif sélectionné. En effet, plus les données sont bruitées, moins le résultat final sera pertinent. Ce pré-traitement se compose de diverse phases : le nettoyage permettant de compléter les données manquantes et d'éliminer les valeurs aberrantes et les doublons, l'homogénéisation conduisant à l'intégration des données provenant des sources diverses, la standardisation amenant à une normalisation et à une mise à l'échelle des données.

ECD-4 : **Transformation :** Les données pré-traitées doivent être transformées afin de les représenter dans un codage adéquat pour l'application de méthodes de fouille de données. Cette étape clée impose parfois de remanier les données pour s'adapter à la phase de fouille.

ECD-5 : **Fouille de données :** La ou les méthodes de fouille de données, traduction de "*data mining*", doivent être choisies en fonction des objectifs visés(classification, segmentation, association) et / ou de leur disponibilité. La complémentarité des techniques amènent avantageusement à les combiner pour un même jeu de données. Insistons ainsi que la fouille de données n'est qu'une étape d'ECD. Malheureusement, il y a souvent une confusion entre ces deux termes, au point que certains auteurs parlent indifféremment de fouille de données et d'extraction de connaissances à partir de bases de cdonnées. Cette étape consiste à investir / appliquer les techniques aux données transformées [Pas00], [ALS03]. Certaines méthodes nécessitent d'asjuster des paramètres et de faire des essais pour optimiser les solutions. Nos travaux de recherche concernent esentiellement la fouille de données. Aussi, nous développons cette phase dans la section qui suit.

ECD-6 : **Interpétation et validation :** Les modèles issus de la fouille de données sont ensuite interprêtés, avec prise en compte de l'avis de l'expert (eu égard à ses problématiques et objectifs). Leur interprétation conduit à la validation ou à la réfutation qui pourrait remettre en cause tout le processus ou une partie du processus d'ECD [LT04]. Diverse méthodes de validation sont envisageables. Les modèles issus d'une classification pourront être vérifiés en

premier lieu par un expert, puis la validation sera complétée par des tests statistiques sur des bases de cas existantes. Pour des techniques d'apprentissage non supervisées telles que la segmentation et l'association, la détermination de la pertinence des modèles obtenus est essentiellement une affaire d'expertise. Pour ce qui est de la classification supervisée, une validation croisée est récommandable à partir de trois ensembles de données, c'est-à-dire un ensemble d'apprentisage, un ensemble de tests et un ensemble de validations. Il est même possible d'effectuer une validation croisée permettant de calculer l'erreur d'un modèle construit sur un ensemble par rapport aux entités de ce même ensemble. On retrouve aussi ce problème de validation en apprentissage automatique.

1.2.3 Les méthodes de fouille de données

L'objectif de la fouille de données est de faire émerger par des méthodes algorithmiques des tendances ou des schémas ("patterns") à partir d'un grand volume de données pré-traitées. La fouille de données peut se faire selon deux orientations différentes(Fayyad et al. 96) :

- dans un but explicatif de la masse des données soumise à l'analyse,
- dans un but de prise de décision par prévision de valeurs inconnues ou de comportements futurs.

La fouille de données a plusieurs tâches : décrire, prédire et / ou d'analyser les liens ou dépendances entre les données [DM00].

- **Description :** Cela consiste à résumer les données en en fournissant une description abstraite. Les techniques employées sont les méthodes de classification non supervisées dont l'objectif est de construire un ensemble de classes à partir d'un ensemble non structuré de données. Ces méthodes s'appellent aussi méthodes de segmentation ou de regroupement. Une telle classification peut être engendrée par des méthodes d'analyse de données (analyses factorielles, classifications hiérarchiques ascendantes ou descendantes, classification pyramidale, nuées dynamiques, plus proches voisins). La classification peut être dirigée. Les hiérarchies conceptuelles apparaissent plus faciles à interprêter que celles basées sur des calculs numériques.
- **Prédiction :** Il s'agit de savoir classer correctement de nouveaux exemples grâce aux règles construites à partir du jeu d'essai. Ces méthodes sont basées sur une classification automatique supervisée. Ici les classes sont déjà prédéfinies par l'utilisateur et décrites par des exemples. Ce sont des techniques d'induction : le système d'induction des règles et les arbres de décision au niveau symbolique, les réseaux de neurones (tels que le perceptron) au niveau numérique. Les méthodes symboliques ont l'avantage de produire des modèles facilement compréhensibles pour un utilisateur et intègrent les connaisances du domaine. Il est loisible d'envisager une méthode de partitionnement préalable pour obtenir les classes initiales.

- **Liens :** Les techniques d'analyse de liens permettent de mettre en évidence les dépendances existant entre les objets donnés, les propriétés ou les objets et les propriétés. Elles déterminent les associations de la forme "Si on a tel ensemble de faits, dans tant pourcentage des cas, on a également tel autre ensemble de faits, dans tant pourcentage des cas". Les méthodes développées

sont les régressions du point de vue numérique, et la recherche des règles d'association ou de motifs fréquents ou de motifs séquentiels du point de vue symbolique. Elles ont principalement un but descriptif ; leur utilisation pour la prédiction est généralement contestée.

Bref, les techniques de la fouille de données sont diverses et touchent plusieurs domaines tels que les statistiques, l'analyse de données, l'algorithmique, l'apprentissage automatique. Quant à la question "Comment choisir une méthode ? ", signalons tout d'abord qu'il n'existe pas de méthode supérieure à toutes les autres parmi les techniques existantes. Les méthodes sont plutôt complémentaires. La sélection d'une méthode se fait en fonction de la nature du problème, mais aussi des données. Les critères du choix sont généralementproposés :

- dans un but explicatif de la masse des données soumise à l'analyse,
- dans un but de prise de décision par prévision de valeurs inconnues ou de comportements futurs de valeurs inconnues.

La fouille de données a plusieurs tâches : décrire, prédire et / ou d'analyser les liens ou dépendances entre les données.

- la tâche que l'on souhaite résoudre,
- l'utilisation finale du modèle obtenu (complexité, performance, pérennité)
- le type de données accessibles,
- les connaissances et compétences disponibles,
- le contexte global.

Dans le présent document de synthèse, nous nous intéressons essentiellement à l'approche symbolique centrée autour de l'implication statistique, ou de l'analyse statistique implicative (ASI), et de la classification (non supervisée) par treillis. L'approche symbolique a l'avantage d'être centrée sur la description des objets (ou entités), en tenant compte de connaissances du domaine, et d'être capable de fournir des explications [Kod99].

Classification par treillis

Rappelons qu'un treillis est un ensemble ordonné dans lequel tout couple d'éléments possède une borne inférieure et une borne supérieure [Ö44], [Eve55], [DP94]. Nous nous intéressons ici au cas où l'ensemble ordonné est fini, donc aux treillis finis [LYKC02], [BW95]. Tout comme un ordre partiel, un treillis peut être représenté par un diagramme de Hasse [Pol98], [Bor86]. Par exemple, la famille des parties d'un ensemble fini ordonné par l'inclusion est un treillis dit treillis de parties [Gu90].

1.3 Contributions

1.3.1 Intensité d'implication et cohésion implicative selon Gras

1.3.1.1 Processus de traitements et valeurs critiques de l'intensité d'implication de Gras

Dans ([Tot94, GSB+96]), nous avons étudié, entre autres, la relation fonctionnelle entre l'indice d'implication de Gras $IndImp_{Gr}$ et l'indice de similarité Sim_{Ler} de Lerman, puis entre l'indice d'implication de Gras et la statistique de Khi-Deux d'indépendance (χ^2).

D'une part, la relation $\frac{\chi^2}{IndImp_{Gr}(X,\overline{Y})^2} = \frac{n^2}{n_X n_Y}$ montre que les deux concepts χ^2 et $IndImp^2$ ne se superposent pas et que les valeurs critiques d'acceptabilité de

l'intensité d'implication ne sont pas constantes mais dynamiques, dans le sens où elles dépendent des trois paramètres à savoir la taille de l'échantillon n, le support n_X de la prémisse X et le support n_Y du conséquent Y ; ce qui permet d'accroître la crédibilité des intensités d'implications observées. Notons qu'ainsi corrélativement à une valeur limite de l'indice d'implication correspond une valeur limite de nombre de contre-exemples $n_{X\overline{Y}}$ au-dessus de laquelle il devient inacceptable de valider l'implication $X \rightarrow Y$ au même seuil. Nous avons pu élaborer les abaques donnant les valeurs critiques gaussiennes pour un niveau de confiance fixé de χ^2. Ce qui a permis d'améliorer considérablement la confiance à l'égard de l'intensité d'implication. La prise en compte de ces valeurs critiques s'avère opérationnellement utile lors de l'utilisation du logiciel associé de Classification Hiérarchique Implicative et Cohésitive ou C.H.I.C.

D'autre part, la relation fonctionnelle $\frac{IndImp_{Gr}(X,\overline{Y}}{Sim_Ler(X,Y)} = -\sqrt{\frac{n_Y}{n_X}}$ montre que le rapport des deux indices ne dépend que de la réalisation de la variable conséquent Y, de la variable concomitante de sa négation et non pas de celle de la prémisse X, et que les deux concepts similarité de Lerman et indice d'implication de Gras sont proches mais différentes, en outre qu'il peut y avoir similarité sans qu'il y ait implication et réciproquement.

Dans [GT95a], à travers l'analyse d'un jeu de données d'une recherche en didactique de la notion de probabilité conditionnelle, nous avons proposé et valorisé une méthodologie de l'analyse statistique implicative exploitant les liens fonctionnelles entre l'indice d'implication de Gras et la similarité de Lerman d'une part, puis entre l'indice d'implication de Gras et le coéfficient de corrélation linéaire d'autre part : il s'agit de l'analyse statistique implicative post-similaritée, puis de l'analyse statistique implicative post-corrélation. Profitant du contexte à population *a priori* hétérogène, nous y avons proposé une méthodologie de recherche et d'exploitation d'un chemin implicatif globalement invariant qui rend compte d'une hiérarchie implicative révélatrice d'une taxonomie interpétable, entre autres, en termes de l'ordre de complexité cognitive. Puis, nous y avons montré une méthodolo- gie (basée sur la visualisation) d'analyse implicative par adjonction, autour du chemin implicatif globalement invariant (obtenu avec le mélange des deux populations homogènes), de deux graphes implicatifs spécifiques des deux populations relativement homogènes. Ainsi, nous avons mis en évidence les deux résultats didactiques suivants :

- l'association de l'arborescence et des conceptions causaliste et chronologiste est source d'obstacle à la réversibilité du concept de probabilité conditionnelle et la commutativité du connecteur logique *et* ;
- mais aussi, la pratique de l'arborescence est efficace dans la résolution d'un problème de type bayésien.

Dans [GT95b], toujours avec un jeu de données de recherche en didactique des mathématiques, nous avons montré que l'analyse statistique implicative révèle des informations supplémentaires et plus décisives comparativement à d'autres méthodes d'analyses statistiques multivariées classiques (i.e. analyse factorielle des correspondances multiples, test d'hypothèses, classification hiérarchique de similarité de Lerman) ainsi qu'à la méthode traditionnelle d'analyse qualitative des données. Ainsi, au sujet de l'apprentissage du concept de probabilité conditionnelle, les résultats obtenus comprennent les suivants :

- confondre l'accroissement de l'information avec celui de la probabilité" implique "adopter la conception causaliste de la probabilité conditionnelle" ;

– le "refus incondtionnel de l'éventuel accroissement de la probabilité" implique "adopter la conception "toujours chronologiste" de la probabilié conditionnelle".

Comme conséquences didactiques de cette analyse statistique implicative du jeu de données cerné, on en conclut entre autres que :

– trop renforcer le caractère causal risque de faire confondre variation d'incertitude et variation de probabilité, sachant que du point de vue formalisme mathématique, il n'y a aucune place pour la relation causale, bien que beaucoup de situations à modéliser la contiennent effectivement ;
– trop insister sur la chronologie risque d'accentuer la confusion entre indépendance stochastique et incompatibilité.

Par suite, en matière de la pédagogie des probabilités conditionnelles, autant on peut considérer que les situations introductives présentant des caractères où causalité ou temps sont des facteurs qui donnent effectivement sens au conditionnement, autant il faut trouver aussi des situations qui déséquilibrent et font évoluer les représentations mentales en découlant avant qu'elles ne s'installent définitivement. Faute de quoi, toute situation formelle ne contenant pas ou causalité ou chronologie ne serait pas résoluble par les apprenants débutants, en particulier tout ce qui est relatif au théorème de Bayes.

Articles :

– 1. R. Gras & A. Totohasina (1995), Conceptions d'élèves sur la notion de probabilité conditionnelle révélées par une méthode d'analyse des données : implication-similarité-corrélation, in Revue Educational Studies in Mathematics 28, Kluwer Academic Publishers, Printed in the Netherlands, 1995, 337-363.
– 2. R. Gras & A. Totohasina (1995), Chronologie et causalité, conceptions sources d'obstacles épistémologiques à la notion de probabilité conditionnelle, in revue Recherche en Didactique des Mathématiques, Vol.15, n¡1, La Pensée Sauvage (édts), Grenoble, France, 1995, 49-95.
– 3. André Totohasina (1994), Notes sur les valeurs critiques de l'indice d'implication de Gras, Rapport technique de recherche, Équipe statistique, Département de mathématiques, Université du Québec à Montréal, Canada (non publié).

Ouvrage collectif :

R. Gras, S. Ag Almouloud, A. Larher, H. Ratsimba, A. Totohasina (1996), *L'Implication statistique . Une nouvelle méthode d'Analyse exploratoire*, La Pensée sauvage (éditions), Grenoble, France, 1996.

1.3.2 Mesures de la qualité des règles d'association et normalisation. Analyse des concepts formelle et correspondance de Galois

1.3.2.1 Propriétés de la mesure M_{GK} et extension

Dans [Tot03], nous avons proposé une normalisation des mesures de qualité des règles d'association respectant les cinq situations de références intuitives en probabilités et statistiques, à savoir l'incompatibilité, la dépendance négative ou répulsion, l'indépendance statistique, la dépendance positive ou l'attraction et l'implication logique. Ce qui a permis de fournir un élément de réponses au souci d'un manque de

vision unificatrice face au foisonnement des mesures de qualité dans la littérature [PS91b], [FPSM92]. Nous avons montré que cette normalisation permet utilement de comparer une grande majorité des mesures de qualité des règles d'association, via la mesure M_{GK} de Guillaume-Kentchaff, qui se trouve normalisée et égale à sa propre normalisée. Ce qui montre que la mesure M_{GK} joue un rôle central au sein des mesures de qualité normalisables et dont la normalisée égale justement M_{GK}. La normalisation proposée est basée sur le résultat d'analyse mathématique élémentaire qui affirme que deux intervalles de $\mathbb{R}$ de même type, sont homéomorphes et difféomorphes. D'autres propriétés mathématiques intéressantes de cette dernière y ont été dégagées. D'où notre intérêt sur cette mesure M_{GK} et sur l'approfondissement de ses propriétés mathématiques dans la suite du présent document de synthèse. Elles ont fait l'objet d'une bonne partie d'une thèse de doctorat élaborée sous ma propre co-direction(Daniel Feno [Fen07].

Dans [TRD03], [TRD04]) et ([TR05]), nous avons proposé une méthode de traitement des données latticielles basées sur la mesure implicative orientée normalisée M_{GK} de Guillaume-Kentchaff, alors appelée *ION* eu égard à sa sémantique d'implication statistique.

Dans [FDT06a], à la lumière d'autres propriétés mathématiques de la mesure M_{GK}, ainsi que par le théorème de condition nécéssaire et suffisante pour qu'une mesure de qualité soit normalisable, nous avons proposé une classification originale des mesures de qualité des règles : les mesures de qualité à normalisée égale à M_{GK}, les mesures normalisables à normalisée différente de M_{GK} et celles non normalisables.

Dans ([DFT06, FDT06b, FDT07, DdRFT07]), à la lumière d'autres propriétés ma-thématiques de la mesure M_{GK}, notamment celle consistant à éviter systématiquement les règles dont prémisse et conséquant sont indépentants ou proches de l'indépendance statistique, contrairement à la mesure *confiance*, intégrant la correspondance de Galois et l'opérateur de fermeture de Galois en théorie des tréllis, nous avons élaboré des axiomes d'inférences correctes, des caractérisations et des algorithmes de génération de bases des règles d'association valides au sens de la mesure M_{GK} dans un contexte binaire. Une base des règles d'association est cette fois constituée par la réunion des quatre bases sectorielles : base des rèrgles positive exacte, base des règles négative exacte, base des règles positives approximatives et base des règles négatives approximatives. Et plus spécifiquement dans [TR05], nous avons montré la capacité de la mesure M_{GK} pour cerner à la fois les rè gles d'association positives et celles négatives, i.e. les règles négatives à droite et celles négatives à gauche.

Dans [DRT07] nous donnons, entre autres, un théorème démontrant la possibilité d'élaborer des abaques des valeurs critiques pour la mesure M_{GK}. En fait, il s'avère que les seuils de signification des valeurs contextualisées de M_{GK} sont dynamiques.

Articles

- 1. Totohasina A. (2003) *Normalisation des mesures probabilistes de la qualité des règles d'association*, in Proceedings Société Statistique de France XXXVè Journées SFDS, Université Lumière Lyon 2, France, 2003, pp. 985-988.

- 2. Totohasina A., Ralambondrainy H., Diatta J. (2004), *Notes sur les me-*

sures probabilistes de la qualité des règles d'association : un algorithme efficace d'extraction des règles d'association implicative, Proceedings Colloque Africain sur la Recherche en Informatique, CARI, Hammamet, Tunisie, 2004, 511-512.

- 3. Totohasina A., Ralambondrainy H. , Diatta J.(2005), *Une vision unificatrice des mesures probabilistes de la qualité des règles d'association booléenes et un algorithme efficace d'extraction des règles d'association implicative*, proceedings of Atelier francophone de Traitement et Analyse de l'Information : Méthodes et Applications, TAIMA 2005, Hammamet, Tunisie, 26 septembre-1er octobre 2005, 375-380.
- 4. Totohasina A., Ralambondrainy H. (2005), *ION : a pertinent new measure for mining information from many types of data*, proceedings of The 2005 International Conference on Signal-Image Technology & Internet- Based Systems (SITIS'05), November 27th - December 2nd 2005, The Hilton Hotel, Yaoundé, Cameroon, 202-207.
- 5. Feno D., Diatta J., Totohasiina A.(2006), *Normalisée d'une mesure probabiliste de la qualité des règles d'association : étude de cas*, Proceedings of EGC 06, Qualité des données et des Connaissances (QDKQ), Lille, France, 2006, 25-30.
- 6. Feno D., Diatta J., Totohasina A.(2006), *Une base pour les règles d'association valides au sens de la mesure de qualité* $\mathrm{M_{GK}}$, Proceedings of XIIIme Rencontres SFC 06, Socié Française de Classification, Metz, France, 2006.
- 7. Feno D., Diatta J., Totohasina A.(2007), *Une base pour les règles d'association valides au sens de la mesure de qualité* $\mathrm{M_{GK}}$, in Revue de la Nouvelle Technologie de l'Information, RNTI, issue spéciale de SFC'2006, version longue, 11 pages (à paraître).
- 8. Feno D., Diatta J., Totohasina A.(2006), *Génération des bases pour les règles d'association* $\mathrm{M_{GK}}$*-valides*, Proceedings of XIVmes Rencontres SFC 07, Socié Française de Classification, TELECOM Paris, 5-7 Septembre 2007, 101-105.
- 9. Feno D., Diatta J., Totohasina A.(2006), *Galois lattices and Bases for* $\mathrm{M_{GK}}$*-valid association rules*, Long papers, in Procceedings of4^{th} International Conference on Concept Lattices and Their Applications, CLA 06, Hamamet, Tunisie, october 30-november 1st, 2006, 127-138.
- 10. Diatta J., Ralambondrainy H., André Totohasina (Janvier 2007), *Towards a unifying Implicative Normalized probabilistic quality measure for association rules*, in Quality Measure in Data Mining, Series Studies in computational intelligence, Guillet Fabrice & Hamilton Howard editorsVol 43, 10^{th} Chapter , 237-238.
- 11. Feno D., Diatta J., Totohasina A.(2007), *Galois lattices and Bases for* $\mathrm{M_{GK}}$*-valid association rules*, Revisited version, in Lecture Note in Artificial Intelligence, Belohlavek & al editors, special issue of CLA 2006 (à paraître).
- 12. Diatta J., Feno D., Totohasina A.(2007), *Mining Bases for* $\mathrm{M_{GK}}$*-valid association rules*, Revisited version, in Global Journal for Pure and Applied Mathematics, GJPAM, Research India Publications, (9 pages)(Accepted to appear).

1.3.3 Traitement des variables quantitatives

Dans [Tot06] et [TRD05], souhaitant éviter trop de perte d'information, nous avons proposé une nouvelle méthode d'extraction des règles d'association quantitative à partir des données quantitatives basée sur les concepts d'espérance et de variance conditionnelles, via le théorème de décomposition de variance, ainsi que sur une mesure de qualité implicative normalisée telle la mesure M_{GK} de Guillaume-Kentchaff [Gui00]. Ladite méthode ne nécessite plus la technique traditionnelle de discrétisation, qui est effectivement à l'origine dune perte d'information considérable, ni les tests statistiques de comparaison des moyennes ou variances. Nous apportons ainsi une solution palliant la faiblesse communément constatée du fait de se contenter d'explorer des données booléennes obtenues par discrétisation souvent arbitraire pour explorer les données quantitatives. Signalons que cette méthode peut très bien s'appliquer en sciences physiques pour résoudre statistiquement un problème d'identification d'une variable ayant une influence plus notable sur une autre variable, à l'instar de [Fém90] et d'un travail sur l'écoulement plan d'un fluide chaud effectué sous mon encadrement par un étudiant de DEA de Mécanique de Fluides et Système Énergétique de la Faculté des Sciences à l'Université d'Antsiranana (Cf. mémoire en cours de PASCAL Petera). Dans le contexte de l'étude des paniers des ménagères, ce problème ouvert est clairement évoqué dans [FPSS96] en les termes suivants : " *We did not consider the quantitative or values of the items bouth in a transaction, which are important for some application. Finding such rules needs further work*".

Articles :

- 1. Totohasina A. (2005), *Une nouvelle méthode d'extraction des règles d'association quantitative*, Proceedings of Atelier francophone de Traitement et Analyse de l'Information : Méthodes et Applications, TAIMA 2005, Session invitée Analyse Symbolique de l'Information ASI, Hammamet, Tunisie, 26 septembre-1er octobre 2005, 54-59.
- 2. Totohasina A.(2006), *Extraction des règles d'association à séman- tique d'implication à partir des données quantitatives*, Proceedings of XXXVIIIemes journées de Société Française de Statistique, SFDS 2006, 29 mai-2 Juin 2006, Clamart, France, 2006.

1.3.4 Classification

Dans ([JHA07]) nous avons proposé une nouvelle méthode formelle, fondée sur la classification, pour rechercher les règles d'association dans un contexte de description ordonnées, c'est-à-dire un contexte où l'espace de description des objets est un ensemble partiellement ordonné. Notre approche consiste à classifier l'ensemble des objets, puis de considérer comme candidat prémisse ou conséquent 'une règle d'association l'intension d'une des classes qui résul- tent de cette classification. Ainsi, un espace de recherche de règles d'association valides est entièrement déterminé par les classes obtenues, et variera selon la mesure dec dissimilarité utilisée, la méthode de classification adoptée, ou la structure de classification construite. L'association des règles à des classes optimisant un critère est un facteur de pertinence qui renforcerait la qualité de ces règles, évaluée par ailleurs en utilisant une ou plusieurs mesures de qualité proposées dans la littérature.

Article :

1. Diatta J., Ralambondrainy H., Totohasina A.(2007), *Règles d'association dans un contexte de descriptions ordonnées*, Proceedings of XIVmes Rencontres SFC 07, Socié Française de Classification, TELECOM Paris, 5-7 Septembre 2007, 94-96.

1.3.5 Articles acceptés pour communications orales aux colloques nationaux

- 1. Totohasina A. (avril 2005), *L'implication statistique orientée normalisée (ION) : vers un outil de la fouille d'un grand volume des données de divers types*, Actes de Forum de Recherches du MENRS, Toamasina, 22 pages, 2005.
- 2. Totohasina A. (2000), *Étude d'une situation didactique en statistique double par l'implication statistique*, Journées de la recherche du MINSUP, Fianarantsoa, 20 pages, Juin 2000.
- 3. Totohasina A. et Feno D. R. (1999), *Suggestions sur la pédagogie de résolution de problèmes. Taxonomie à partir de l'implication statistique de Gras*, Actes du colloque international sur la didactique des disciplines, ENS d'Antanànarivo, 12 pages, 1999.

1.3.6 Articles soumis à un colloque international(fév. 2008)

- 1. Totohasina A. (2008), *De la qualité des règles d'association. Une normalisation unificatrice des mesures, rôle de* M_{GK}, 8 pages.(soumis).

  Résumé : Dans le double objectif de comparaison et d'une vision unificatrice des différentes mesures de qualité de règle d'association, nous proposons une définition d'une mesure probabiliste normalisée de qualité et une *normalisation* d'une mesures non normalisée. Nous montrons qu'il est possible de construire une infinité de mesures normalisées satisfaisant à la fois les cinq situations de référence plus ou moins intuitives telles l'incompatibilité, l'indépendance statistique, la nature de dépendance et l'implication logique. Cependant notre approche diffère de celle présentée dans [HH99, AZ04]. Elle est en fait guidée par les propriétés des intervalles réels, car il est claire que ces mesures de qualité prennent leurs valeurs soit dans un intervalle borné, soit dans un intervalle ouvert non borné.

- 2. Totohasina A., Feno D. R.(2008), *De la qualité des règles d'association : Étude comparative des mesures* $\mathrm{M_{GK}}$ *et* $Confiance$. (8 pages) (soumis).

  Résumé : Dans le présent papier, nous donnons des propriétés mathématiques intéressantes de $\mathrm{M_{GK}}$ et un algorithme de génération de base composite de règles d'association valides selon la mesure $\mathrm{M_{GK}}$ de Guillaume-Kentchaff, avec un exemple d'application sur des donnée concrètes à l'appui. Parallèlement à cela, il permet de comparer $\mathrm{M_{GK}}$ et la mesure pionnière $Confiance$ en fouille de données quant à la pertinence des règles produites.

Chapitre 2

Règle d'Association. Mesure Probabiliste de Qualité

2.1 Règle d'association dans un contexte binaire. Mesure probabiliste de qualité

2.1.1 Base de données ou contexte formel : treillis, motifs, règle d'association

Nous nous plaçons tout d'abord dans le cadre d'un contexte de la fouille de données binaires $\mathbb{K} = (\mathcal{O}, \mathcal{A}, \mathcal{R})$, où $\mathcal{O}$ est un ensemble fini d'entités ou d'objets, $\mathcal{A}$ un ensemble fini non vide d' attributs (ou de variables, ou items, ou des propriétés) et $\mathcal{R}$ une relation binaire de $\mathcal{O}$ vers $\mathcal{A}$. Comme nous le reverrons plus bas, le problème de la fouille des règles d'association peut être entièrement traité dans le cadre de l'Analyse Formelle de Concepts (A.F.C.). L'A.F.C. fournit un cadre théorique fondamental pour plusieurs algorithmes de fouille des règles d'association. En A.F.C., la base de données $\mathbb{K} = (\mathcal{O}, \mathcal{A}, \mathcal{R})$ est appelée un contexte formel ([Wil82] ; [GW99] ;[BM70]).

Parfois on assimile la relation binaire $\mathcal{R}$ à son graphe. L'appartenance d'un couple $(o, a) \in \mathcal{O} \times \mathcal{A}$ au graphe de la relation $\mathcal{R}$ exprime le fait que l'objet o possède l'attribut ou la propriété a. Signalons qu'ainsi tout attribut peut être identifié à une application de $\mathcal{O}$ dans le doubleton $\{0, 1\}$, où la valeur 1 mesure la présence de l'attribut chez un objet de $\mathcal{O}$. Dans toute la suite, n désigne la cardinalité de $\mathcal{O}$ ($n = |\mathcal{O}|$).

Tout sous-ensemble X de $\mathcal{A}$ s'appelle un *motif* ou ou *transaction* ou ou *itemset* de $\mathcal{A}$, et sa négation logique notée $\overline{X}$ le *motif négatif associé au motif* X, et tout élément de $\mathcal{O}$ un *objet* ou une *entité* de $\mathcal{O}$. Pour tout motif X de $\mathcal{A}$, remarquons les huit points suivants :

(a) $\forall a \in \mathcal{A}$, $1 - a$ =le complément à 1 de a, i.e. $(1 - a)$ identifie l'absence de l'attribut a chez une entité.

(b) $\forall e \in \mathcal{O}, X(e) = 1 \iff \forall a \in X, e\mathcal{R}a, i.e. a(e) = 1$; soit $X = \wedge_{a \in X} a =$ la conjonction des présences d'un nombre fini d'attributs de $\mathcal{A}$.

(c) $\overline{X} = \bigvee_{1-a} a \in X =$ la disjonction des absences des attributs de X.

(d) $X' = \{e \in \mathcal{O} | \forall x \in X, e\mathcal{R}x\}$, *i.e.* l'ensemble de toutes les entités communes à tous les éléments de X : c'est le dual d'un motif X de $\mathcal{A}$, ou l'extension du motif X.

(e) De façon duale, pour une partie $E \in \mathcal{O}$, le motif contenu par E, noté et défini par $E' = \{a \in \mathcal{O} / \forall e \in E, a(e) = 1\}$, est appelé l'intension de E : c'est l'ensemble des attributs communs aux objets de E.

(f) $\overline{X}' = \mathcal{O} - X' = \overline{X'}$; cette coïncidence renforce l'appellation de motif négatif pour $\overline{X}$.

(g) $\overline{\overline{X}} = X$: on retrouve ainsi le caractère involutif de l'action de négation en logique formelle, i.e. la loi de De Morgan.

(h) $X \subseteq \mathcal{A}$, mais $\overline{X} \nsubseteq \mathcal{A}$.

Il est facile de voir que pour deux motifs X et Y du contexte, en termes d'extension on a : $(X \wedge Y)' = X' \cap Y'$ et $(X \vee Y)' = X' \cup Y'$. Il existe plusieurs modélisations probabilistes de la fouille de données dans un contexte binaire (voir par exemples [LGR81], [Ler81], [Ler84]. Dans notre modélisation, qui reste en fait souvent implicite dans la littérature, puisqu'a priori on n'a pas de raison pour le réfuter, nous nous plaçons dans le cadre d'une hypothèse d'équiprobabilité des événements élémentaires de $\mathcal{O}$. D'où la considération de ll'espace probabilisé discret $(\mathcal{O}, \mathcal{P}(\mathcal{O}), P)$, P étant la probabilité uniforme. Par conséquent, pour tout X de $\mathcal{A}$, en notant $n_X = |X'|$ la cardinalité de X', $Supp(X) = \frac{n_X}{n}$ représente une estimation de la probabilité $P(X')$ de l'événement X'. De plus, par glissement de terminologie qui se justifié par la dualité entre extension et intension, il s'avère naturel d'adopter les définitions suivantes : Deux motifs sont indépendants (resp.dépendants), si leurs extensions respectifs sont indépendants (resp. dépendants) dans l'espace probabilisé $(\mathcal{O}, \mathcal{P}(\mathcal{O}), P)$. La littérature atteste que ceci a déjà été longtemps intuitivement accepté.

Dans la suite, pour un motif X de $\mathcal{A}$, nous appellerons $\overline{X}$ le motif *n*égatif associé, tandis que X un motif *p*ositif. Le mot *motif* désignera indifféremment un *motif positif* ou un *motif négatif*.

Le Tableau 2.1 présente un contexte binaire $\mathbb{K} = (\mathcal{O}, \mathcal{A}, \mathcal{R})$, où $\mathcal{O} = \{e_1, e_2, e_3, e_4, e_5\}$ et $\mathcal{A} = \{A, B, C, D, E\}$. Pour $X = \{B, C\}$,

on a $X' = \{e_2, e_3, e_5\}$ et $\overline{X}' = \{e_1, e_4\}$, $Supp(X) = \frac{1}{2} = 1 - Supp(\overline{X}')$.

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

TAB. 2.1 – Contexte binaire

Dans la littérature, les règles d'association considérées sont généralement des règles utilisant essentiellement les motifs positifs. Certaines applications nécessitent non seulement la découverte des règles d'association de la forme "si X, alors Y", mais aussi celle des règles de la forme "si X, alors $nonY$" ou "si $nonX$, alors Y". Ainsi, nous introduisons une définition plus générale des règles d'association utilisant les motifs positifs et ceux négatifs. Nous adoptons donc la définition suivante.

Définition 1 *(a) Une règle d'association d'un contexte binaire* $\mathbb{K}$ *est un couple* (U, V)*, noté* $U \rightarrow V$*, où* U *et* V *sont des motifs positifs ou négatifs et* $V \neq \emptyset$ *(si* V *est un motif positif),* $V \neq \overline{\mathcal{A}}$ *(si* V *est un motif négatif) : on lit communément "si* U*, alors* V*". Comme son nom l'indique, elle exprime une association ou un lien orienté entre* X *et* Y*. Parfois on l'appelle aussi une* quasi-implication [LGR81].

(b) Pour une règle d'association $U \rightarrow V$, U et V sont appelés respectivement la *prémisse* et le *conséquent* de la règle.

(c) Pour toute règle d'association $U \rightarrow V$, la proportion des entités possédant le motif $U \cup V$, notée n_{UV}/n est le support de $U \rightarrow V$, soit $\text{supp}(U \rightarrow V) = n_{UV}/n = P(U' \cap V')$, et la proportion des entités contenant le *conséquent* V parmi ceux possédant la *prémisse* U mesure la *Confiance* de la règle $U \rightarrow V$, on note : $\text{conf}(U \rightarrow V) = n_{UV}/n_U$, c'est une estimation de la probabilité conditionnelle sachant U' de V', i.e. $P(V'/U')$.

(d) Ces deux critères *Support* et $Confiance$ sont ainsi appelés deux mesures objectives de la qualité ou mesures objectives d'intérêt d'une règle d'association.
En fait la littérature atteste qu'il existe actuellement une quarantaine de mesures objectives de la qualité des règles d'association ([GH06, HGB05b]) pour diverse raisons.

Remarque 1 *En général, on requiert d'une part que le conséquent d'une régle d'association ne soit pas vide et, d'autre part, que la prémisse et le conséquent soient disjoints, pour ne pas considérer des règles triviales qui n'apportent à l'utilisateur aucune information utile.*

Toutefois, malgré ces restrictions, le nombre de régles d'association conformes à la définition ci-dessus reste très élevé. Ainsi, dans un souci d'informativité et de pertinence, on n'en retient que ceux qui sont valides au sens d'une (un ensemble de) mesure(s) de qualité. La plupart de ces mesures de qualité sont probabilistes (*i.e.* se définissent entièrement à partir d'un tableau de contingence) et les plus connues d'entre elles sont le *support*, et la $confiance$. Ainsi, contrairement à certains auteurs, dans notre approche il n'y a pas lieu de distinguer entre mesure statistique, mesure descriptive et mesure probabiliste. Dans la pratique, on fixe au préalable un minimum de seuil de support noté $minsupp$ et un seuil minimum de confiance noté $minconf$:
Une règle d'association $U \rightarrow V$ est alors valide au sens de *support-confiance*, si $\text{supp}(U \rightarrow V) \geq minsupp$ et $\text{Conf}(U \rightarrow V) \geq minconf$.

En effet, s'il est encore vrai que les deux mesures de la qualité *Support* et $Confiance$ sont les plus utilisées, elles produisent de réels problèmes [LMVP03], [Lal02], [LT04]. D'abord, les algorithmes utilisés pour générer les règles d'association d'un contexte binaire engendrent un très grand nombre de règles qui sont très difficiles à gérer et dont beaucoup n'ont que peu d'intérêt.

Ensuite, adopter la condition de support comme le moteur du processus d'extraction écarte les règles ayant un petit support, alors que certaines peuvent avoir une très forte Confiance et peuvent présenter un réel intérêt . Enfin, l'utilisation exclusive des mesures de qualité objectives support et confiance ne suffit pas pour garantir la qualité des règles détectées, car contrairement à la requête de dépendance,

elle produit facilement des règles dont la prémisse et le conséquent sont statistiquement indépendants ou proches de l'indépendance, malgré une confiance élevée : car à tort, une règle à conséquent de forte probabilité serait alors sélectionnée valide, même s'il n'y a pas de lien statistique entre le conséquent et la prémisse.

En effet, comme le montre l'exemple du Tableau 2.1.1 (extrait de [LT04]), la règle $X \rightarrow Y$ possède un Support élevé (si l'on considère $n = 100$) et une Confiance élevée : $\text{supp}(X \rightarrow Y) = 72$ et, $\text{conf}(X \rightarrow Y) = 90$. Cependant, la Confiance de cette règle est égale à la probabilité $p(Y')$, soit $P(Y'|X') = P(Y')$; ce qui est la définition de l'indépendance statistique de X et Y et n'apporte aucune information nouvelle. Nous donnerons plus bas (voir 5 la définition générale d'une mesure objective,

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

TAB. 2.2 – Faiblesse de l'approche Support-Confiance

i.e. indépendante de la connaissance des données, de la qualié de règle que nous proposons.

Pour deux motifs positifs X et Y du contexte $\mathbb{K} = (\mathcal{O}, \mathcal{A}, \mathcal{R})$, on distingue ainsi quatre types de règles d'association possibles, à savoir :

Définition 2 *(a) Une règle dite* positive *de la forme* $X \rightarrow Y$ *ou* $Y \rightarrow X$ *;*
(b) Une règle dite négative à droite *de la forme* $X \rightarrow \overline{Y}$ *ou* $Y \rightarrow \overline{X}$ *;*
(c) Une règle dite négative à gauche *de la forme* $\overline{X} \rightarrow Y$ *ou* $\overline{Y} \rightarrow X$ *;*
(d) Une règle dite bilatéralement négative *de la forme* $\overline{X} \rightarrow \overline{Y}$ *ou* $\overline{Y} \rightarrow \overline{X}$.

2.1.2 Extraction des règles d'association et Classification par treillis de Galois.

Rappelons brièvement les notions de fermeture, de correspondance de Galois et de treillis en général avant de revenir sur notre contexte de fouille de données. Étant donné un ensemble non vide E, un *opérateur de fermeture* ou tout simplement *une fermeture* sur E est une application ϕ définie sur $\mathcal{P}(E)$ vérifiant les trois conditions suivantes :

(F1) pour tous $A, B \subseteq E$, $A \subseteq B$ *implique* $\phi(A) \subseteq \phi(B)$ (isotonie) ;
(F2) pour tout $A \subseteq E$, $\phi\phi(A) = \phi(A)$ (idempotence) ;
(F3) pour tout $A \subseteq E$, $A \subseteq \phi(A)$ (extensivité).

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

D'une manière générale, pour deux ensembles ordonnés $(E, \leq)$ et $(F, \leq)$, on appelle *correspondance de Galois* entre E et F tout couple d'applications (f, g), avec $f : E \rightarrow F$ et $g : F \rightarrow E$ deux applications telles que pour tous $x, x' \in E, y, y' \in F$, les trois conditions suivantes sont vérifiées :

(G1) $x \leq x'$ implique $f(x) \geq f(x')$ (antitonie) ;
(G2) $y \leq y'$ implique $g(y) \geq g(y')$ (antitonie) ;
(G3) $x \leq g \circ f(x)$ *et* $y \leq f \circ g(y)$ (extensivité).

Dans ce cas de correspondance de Galois, l' application composée $\varphi = f \circ g$(resp $\varphi' = g \circ f$) est l'opérateur de fermeture dans l' ensemble ordonné $(E, \leq)$ (resp. $(F, \leq)$).

Et de façon intuitive, par opposition à une définition algébrique donnée dans [CM00], [CR93], un treillis est un ensemble ordonné dont tout couple d'éléments admet un infimum et un supremum. On parle alors de treillis de Galois lorsqu'on a affaire à deux treillis auxquels est associé une correspondance de Galois. Des riches propriétés de caractérisation de toutes ces notions sont trouvables dans ([BM70], [Mon03], [Mor62], [Ise51],[GW99], [Day92], [Cas99], [CM03], [Dia05]). Pour ce qui est des systèmes implicatifs, on peut voir par exemple [Dom02, DL04], [GD86].

Reconsidérons notre contexte formel $\mathbb{K} = (\mathcal{O}, \mathcal{A}, \mathcal{R})$, où $\mathcal{O}$ et $\mathcal{A}$ sont des ensembles finis. La relation binaire $\mathcal{R}$ induit une correspondance de Galois entre les ensembles ordonnés $(\mathcal{P}(\mathcal{O}), \subseteq)$ et $(\mathcal{P}(\mathcal{A}), \subseteq)$ par le biais des deux fonctions duales f et g définies de la façon suivante :

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

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

Ainsi, l'application $\varphi = f \circ g$ (resp. $\varphi' = g \circ f$) est l'opérateur de fermeture dans l' ensemble ordonné $(\mathcal{P}(\mathcal{A}), \subseteq)$ (resp. $(\mathcal{P}(\mathcal{O}), \subseteq)$), i.e. φ est à la fois monotone croissante (ou isotone) et extensive (i.e. $\forall X \in \mathcal{P}(\mathcal{A})$, on a $X \subseteq \varphi(X)$ au sens de l'ordre inclusion(resp.φ' dans $\mathcal{P}(\mathcal{O})$), et idempotente, *i.e.* $\varphi \circ \varphi = \varphi$ (resp. $\varphi' \circ \varphi' = \varphi'$).

Un motif X de $\mathcal{A}$ est dit fermé (pour l'opérateur de fermeture φ) s'il est égal à sa fermeture, i.e. si $X = \varphi(X)$, *i.e.* si $X = Intension(Extension(X)) = (X')'$. Tout couple de type $(B, X) \in \mathcal{P}(\mathcal{O} \times \mathcal{P}(\mathcal{A})$ tel que $B = X'$ et $X = B'$ est appelé *concept formel* du contexte formel $\mathbb{K} = (\mathcal{O}, \mathcal{A}, \mathcal{R})$. Notons que pour tout motif X, $\varphi(X)$ est un fermé qui peut être différent de X. Néanmoins, en matière de support on a :

Lemme 1 *[PBTL99] Pour tout motif X, on a :* $\text{supp}(\varphi(X)) = \text{supp}(X)$.

Ainsi, le support d'un motif est égal au support de sa fermeture.

Le Lemme 7 est une caractérisation des opérateurs de fermeture utilisant une propriété dite d'indépendance de chemins [Plo73].

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

Dans la suite du document, nous utiliserons la correspondance de Galois (f, g) ainsi définie sur $\mathcal{P}(\mathcal{O})$ et $\mathcal{P}(\mathcal{A})$: c'est la correspondance de Galois du contexte de fouille de données $\mathbb{K} = (\mathcal{O}, \mathcal{A}, \mathcal{R})$. Ainsi, l'analyse des concepts formelle fournit un cadre théorique fondamental pour la fouille des règles d'association. Rappelons que $(\mathcal{P}(\mathcal{O}), \subseteq)$ et $(\mathcal{P}(\mathcal{A}), \subseteq)$ sont des treillis, c'est-à-dire tout couple (X, Y) de $\mathcal{P}(\mathcal{A})$ admet un supremum $X \vee Y$ (disjonction des deux motifs) et un infimum $X \wedge Y$(conjonction des deux motifs), de même pour $\mathcal{P}(\mathcal{O})$: les deux treillis $\mathcal{P}(\mathcal{A})$ et

$\mathcal{P}(\mathcal{O})$ mis en correspondance par la correspondance de Galois (f, g) constituent un treillis de Galois (ou treillis de concepts) associé au contexte formel $\mathbb{K} = (\mathcal{O}, \mathcal{A}, \mathcal{R})$. Ainsi un motif d'un contexte peut être consiédéré comme la conjonction de ses propres attributs. Donc une règle d'association peut s'exprimer comme une quasi-implication d'une conjonction d'attributs sur une autre conjonction d'attributs de type :

$$R : a_1 \wedge a_2 \wedge a_3 \rightarrow a_4 \wedge a_5 \wedge a_6 \wedge a_7.$$

Un ensemble ordonné fini pouvant se représenter graphiquement par son diagramme de Hasse, la construction d'un treillis de Galois permet de regrouper (donc de classifier) conjointement les groupes d'objets et d'attributs en terme de concepts formels (symboliquement en couples des duaux (Intension(Extension), Extension(Intension)).

Remarque 2 *Lorsque l'ensemble fini $\mathcal{A}$ est muni d'une relation d'ordre (notons la $\prec$), il est loisible de le représenter sous forme d'un diagramme de Hasse. Pour cela on procède ainsi : à chaque élément a de $\mathcal{A}$ on fait correspondre un point du plan euclidien ; pour chaque couple (a, b) de $\mathcal{A}^2$ tel que $a < b$ et $\nexists c \in \mathcal{A}$ tel que $a < c < b$, on relie les deux points du plan correspondant ; les conventions du dessin sont les suivantes : si $a < b$, le point correspondant à a est au-dessous du point représentant b. On n'admet pas l'intersection entre la représentation de c et l'ensemble des minorants de $\{a, b\}$. On élimine ainsi du diagramme de Hasse les arcs de transitivité.*

Cette technique de classification par treillis de concepts formels a l'avantage d'être objective (elle ne dépend que de la condition de fermeture) et unique (peu importe l'algorithme de construction) [Val99]. En plus d'autres algorithmes déjà évoqués, sans entrer dans le détail, on peut aussi extraire les règles d'association depuis le treillis de Galois associé au contexte de fouille de donnée [STB$^+$01]. Dans une section qui suit, nous procédons ainsi dans notre étude comparative des mesures de qualié de règles.

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

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

Définition 3 *Un motif X est dit* maximal fréquent *s'il est fréquent et que tous ses sur-ensembles sont infréquents. Formellement, l'ensemble M_F des motifs maximaux fréquents d'un contexte $\mathbb{K}$ est défini par :*

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

Définition 4 – *Un motif X est dit φ-*fermé*, ou tout simplement* fermé*, si $\varphi(X) = X$. Il est dit φ-*fermé fréquent *s' il est à la fois φ-fermé et fréquent. Formellement, l'ensemble des motifs fermés fréquents d'un contexte $\mathbb{K}$ est défini par*

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

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

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

Remarque 1 *Pour un motif $X \subseteq \mathcal{A}$, $\varphi(X)$ sera appelé la* fermeture *de X. Elle correspond au plus petit fermé qui contient X.*

Proposition 1 *[PBTL99] Pour un motif X, le Support de X est égal au Support de sa fermeture, i.e.,*

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

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

Proposition 2 *[PBTL99] Les ensembles M_F des motifs maximaux fréquents et $M\mathcal{F}_F$ des motifs fermés maximaux fréquents sont identiques, i.e., $M_F = M\mathcal{F}_F$.*

Thèorème 1 *Soient $\mathbb{K}$ un contexte de la fouille de données et* minsupp *un seuil minimum de Support. L'ensemble $\mathcal{F}_F$ des motifs fermés fréquents est une représentation condensée des motifs fréquents, i.e., c'est un sous-ensemble de motifs fréquents à partir duquel on peut dériver tous les motifs fréquents et leurs Supports.*

Plusieurs algorithmes de la fouille des règles d'association sont disponibles dans la littérature. L'algorithme nommé APRIORI dévéloppé par Agrawal et srikant ([AIS93] , [AS94]) figure parmi les toutes premières méthodes d'extraction des motifs fréquents en vue de l'extraction des règles d'association intéressantes. Mais, beaucoup de motifs candiadats y demeurent comptés sans que cela soit nécessaire.

Depuis, d'autres algorithmes plus efficaces ont été élaborés (cf. MaxMiner qui partent des motifs fréquents maximaux, CLOSE ([PBTL99], Bastide et al. 2000 [BTP$^+$02]), CLOSET [PHM00], CLOSET [PHM00], CHARM [ZH99], PASCAL (Bastide et al. 2002 [BTP$^+$02]), TITANIC (Stumme et al. 2002 [STB$^+$02]). Or il se trouve que l'ensemble global des motifs fréquents peut être déduit des fermés sans nécessiter d'accéder à la base de données.

L'algorithme CLOSE extrait les motifs fermés fréquents. Alors que PASCAL génère les motifs équivalents, *i.e.* motifs ayant la même extension, donc de même support.
Comme dans l'algorithme pionnier APRIORI, les motifs sont généralement ordonnés par un ordre lexicographique.

2.1.3 Raisonnement sur les règles d'association

Rappelons que la signification intuitive d'une règle d'association $X \rightarrow Y$ est la suivante : "Chaque fois que le motif X apparaît, le motif Y aussi, avec un certain degré d'assurance", ou encore " tout objet qui possède le motif X a tendance à posséder aussi le motif Y, avec un degré de confiance estimée".

Aussi est-il légitime de formaliser le raisonnement voire l'inférence sur les règles d'association. Dans ce sens, par exemple, les travaux de Bastide et al (2000) et de Duquenne et al. (1986) ont pointé la notion de redondance sur ces dernières à cause de la transitivité. Les principes du raisonnement sur les règles d'association sont résumés par les quelques lignes qui suivent [ZO98] [ZO98].

Une règle d'association est $support - confiance$-intéressante (ou valide) si son support et sa confiance sont respectivement supérieurs ou égaux aux seuils respectifs fixés. Une règle dont la confiance vaut 1 est dite dite une règle exacte.

Théoriquement, le raisonnement sur les règles d'association se base sur les les axiomes d'inférence d'Armstrong [Arm74] définis ci-dessous, notamment pour dériver toutes les règles (Support,Confiance)-valides :

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

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

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

Ils sont étendus par les résultats suivants qui donnent des conditions suffisantes pour les règles non exactes [Luo06].

Proposition 3 *Augmentation-à-gauche : Si une règle $X \rightarrow Y$ est exacte, alors pour tout motif Z, la règle $Z \vee X \rightarrow Y$ est exacte. De plus, si le motif $Z \vee X$ est intéressant, alors la règle $Z \vee X \rightarrow Y$ l'est aussi, avec le même support. Pour trois motifs X, Y, Z , si $X \rightarrow Y \vee Z$ est intéressante, alors $X \vee Z \rightarrow Y \vee Z$ l'est aussi (faible augmentation-à-gauche ou FAG). Il en résulte que si un motif intéressant X est tel que* $\text{supp}(X) \geq minconf$, *alors pour tout sous-motif Z de X, la règle $Z \rightarrow X$ est intéres-sante.*

Addition-à-gauche : *Si les règles $X \rightarrow Y$ et $Z \rightarrow Y$ sont exactes, alors $X \vee Z \rightarrow Y$ est exacte ; et si $X \vee Z$ est motif intéressant, alors la règle $X \vee Z \rightarrow Y$ est intéressante. De plus, si $X \rightarrow Y$ et $Z \rightarrow Y$ sont des règles intéressantes telles que le motif conjoint $(X \vee Y \vee Z)$ est intéressant et la règle $Y \rightarrow X \vee Z$ est exacte, alors la règle $X \vee Z \rightarrow Y$ est intéressante.*

Addition-à-droite : *Si $X \rightarrow Y$ et $X \rightarrow Z$ sont des règles exactes, alors la règle si $X \rightarrow Y \vee Z$. De plus, si le motif X est intéressant, alors $X \rightarrow Y \vee Z$ est intéressante et de même support que X. si $X \rightarrow Y$ est exacte et si $X \rightarrow Z$ est intéressante, alors si $X \rightarrow Y \vee Z$ est intéressante, de mêmes support et confiance que la règle si $X \rightarrow Z$. Si $X \rightarrow Y$ est intéressante et $Y \rightarrow Z$ exacte, alors $X \rightarrow Z \vee Y$ est intéressante, de mêmes support et confiance que la règle $X \rightarrow Y$.*

Décomposition : *Si $X \rightarrow Y \vee Z$ est une règle intéressante, alors $X \rightarrow Y$ et $X \rightarrow Z$ sont intéressantes. Cette décomposition demeure même pour les règles d'association approximatives (support, confiance)-valides.*

Transitivité : *Elle concerne seulement les règles exactes. Soit : si $X \rightarrow Y$ et $Y \rightarrow Z$ sont deux règles exactes, alors $X \rightarrow Y$ l'est aussi.*

Triangularisation : *Si X, Y et Z sont trois motifs tels que $X \subseteq Y \subseteq Z$ et* $\text{conf}(X \rightarrow Y) = c_1$, $\text{conf}(Y \rightarrow Z) = c_2$ *et* $\text{conf}(X \rightarrow Z) = c_3$, *alors si $c_3 \geq minconf$, on a : $c_1 \geq minconf$ et $c_2 \geq minconf$. Il s'en suit que pour trois motifs emboîtés $X \subseteq Y \subseteq Z$, si la règle $(X \rightarrow Z)$ est intéressante, alors les 2 règles $X \rightarrow Y$ et $Y \rightarrow Z$ le sont aussi.*

À partir de ces propriétés, on peut construire un ensemble minimal de règles intéressantes depuis lequel on peut dériver les autres règles intéressantes correspondant à des seuils fixés de support et confiance : c'est ce qu'on appelle une base représentative des règles d'association intéressantes pour un contexte formel donné. Elle est unique pour cette contrainte. La dérivation des autres règles peut s'élaborer comme indiqué dans ce qui suit : le système d'inférence qui repose sur la faible Augmentation-à gauche (FAG) et sur la décomposition est appelé le système-LD d'inférence (ou de dérivation) sur les règles.

On trouve des algorithmes de génération de bases des règles d'association (support-confiance)-valides sur les motifs fermés fréquents dans 1 et [Pas00]. Des algorithmes de génération des motifs fermés existent aussi (voir CLOSE [PBTL99], CLOSET [PHM00], CHARM [ZH99], TITANIC [STB+02], PRINCE [HYS05]. Dans [FDT06b, DFT06], nous proposons des axiomes d'inférence pour générer des bases des règles d'association non redondantes et plus cohérentes à partir d'une mesure de qualité plus pertinente. Une synthèse de ces travaux est présentée dans la dernière section 3.4 du chapitre qui suit.

2.2 Mesure probabiliste de qualité (MPQ)

2.2.1 Définitions

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

Définition 5 *Une* mesure de qualité *ou* mesure d'intérêt *des règles est une fonction μ de l'ensemble des règles d'association à valeurs dans $\mathbb{R}$, telle que pour toute règle d'association $U \rightarrow V$, la valeur $\mu(U \rightarrow V)$ dépend exclusivement des quatre paramètres $n, P(U'), P(V')$ et $P(U' \cap V')$, où P désigne la probabilité discrète uniforme sur l'espace probabilisable $(\mathcal{O}, \mathcal{P}(\mathcal{O}))$.*

Notons ainsi que $\mu(U \rightarrow V)$ est entièrement déterminée par le tableau de contingence obtenu en croisant U et V. Remarquons que cette définion est bien justifiée par le passage par l'ensemble des couples d'extensions possibles $\mathcal{P}(\mathcal{O})^2$, permettant d'investir les probabilités, et par le fait que l'extension d'un motif négatif coïncide avec le complémentaire de l'extension du motif positif associé.

L'application μ s'obtient ainsi via la tribu des parties $\mathcal{P}(\mathcal{O})$ de l'ensemble des objets $\mathcal{O}$ (i.e. par la composition faisant intervenir l'Extension et une application réelle définie sur le tableau de contingence correspondant).

Par ailleurs, ne serait-ce qu'à défaut d'additivité, le terme *mesure* ici n'est pas à prendre au sens de la théorie des mesures en analyse mathématique, mais mesure signifie ici un indice qui permet d'évaluer le degré de lien (orienté) entre deux motifs. Aussi, serait-il plus approprié d'utiliser la termilogie d'indice de degré de dépendance dans une règle d'association probabiliste ? Poursuivant le langage largement utilisé dans la littérature, une mesure objective de la qualité des règles est alors explicitée en fonction des critères ou de la sémantique fixés ou souhaités par le chercheur.

Remarque 2 *Effectivement dans ce cadre de contexte binaire, une telle mesure de qualité, dite une mesure probabiliste de la qualité (MPQ) à notre sens, est une fonction qui est complètement déterminée par le tableau de contingence obtenu en croisant les deux motifs en jeu, i.e. la prémisse U et le conséquent V soit K_{UV} (voir Table 2.3). Lorsque les effectifs marginaux n, n_U et n_V sont fixés, la connaissance d'une valeur de la table de contingence, par exemple le nombre de contre-exemples $n_{U\overline{V}}$, détermine les trois autres valeurs. Ainsi les probabilités $P(U')$, $P(V')$ et $P(U' \cap \overline{V'})$ sont obtenues à l'aide des entrées de la table de contingence. Dans toute la suite du document il s'agit des MPQs.*

TAB. 2.3 – Le tableau de contingence K_{UV}

$U \diagdown V$	V	$\overline{V}$	
U	n_{UV}	$n_{U\overline{V}}$	n_U
$\overline{U}$	$n_{\overline{U}V}$	$n_{\overline{U}\overline{V}}$	$n_{\overline{U}}$
	n_V	$n_{\overline{V}}$	n

Sur la stabilté d'une mesure de qualité suite à une perturbation d'un paramètre

Il résulte de cette remarque qu'une mesure de qualitéμ définie sur le contexte binaire $\mathbb{K} = (\mathcal{O}, \mathcal{A}, \mathcal{R})$ se détermine ou se construise par les quatre paramétres associés à une règle d'association : la *taille n de l'échantillon* $\mathcal{O}$*, la cardinalité* n_U *de la prémisse, la cardinalité* n_V *du conséquent et le nombre de contre-exemples* $n_{U\overline{V}}$ *, c'est-à-dire pour toute règle* $U \rightarrow V$*,* $\mu(U \rightarrow V) = f(n, P(U), P(V), P(U \cap \overline{V})$*, ce qui peut se ramener à* $\mu(U \rightarrow V) = \varphi(n, n_U, n_V, n_{U\overline{V}})$*.*

Ainsi, comme on peut toujours supposer qu'il s'agit là de la restriction d'une fonction réelle de quatre variables réelles, l'étude de la stabilité ou de la sensibilité aux perturbations de l'un ou l'autre de ces paramètres peut s'effectuer par la considération de la dérivée partielle correspondante, voire du gradient de la fonction si cela concerne les quatre paramètres à la fois.

Par exemple pour n fixé, et à marginales n_U *et* n_V *fixées, l'analyse de la dérivée partielle* $\frac{\partial f}{\partial P(U \cap \overline{V})}$ *ou* $\frac{\partial \varphi}{\partial n_{U\overline{V}}}$ *: si cette dérivée partielle est négative, alors* μ *décroît au fur et à mesure que le nombre de contre-exemples augmente, ce que l'on souhaite [GDGB07]. En fait, c'est la dérivée partielle de* μ *par rapport à la taille qui puisse renseigner vraiment sur la stabilité d'une règle en passant d'un volume de données à un autre, car cela répond bien à la question telle : une règle* μ*-valide pour une taille n, le reste-t-elle encore pour* n' *légèrement supérieur (ou inférieur) à n ?*

Dans la suite, nous considérons uniquement les mesures de qualité ainsi qualifiées de probabilistes qui demeurent en fait les plus utilisées selon la littérature. Nous en fournissons quelques exemples dans le tableau 2.1.2 donné ci-après. Par souci de la logique formelle (où deux implications contraposées ont la même valeur logique), souhaitant la sémantique d'implication logique (ou d'inclusion en terme d'extensions), nous posons la définition suivante.

Définition 6 *Une mesure de qualité* μ *sera dite* implicative *si pour toute règle d'association* $X \rightarrow Y$*, on a :* $\mu(\overline{Y} \rightarrow \overline{X}) = \mu(X \rightarrow Y)$ *[Tot03].*

Par exemples, les mesures Confiance, conviction, Intensité d'implication, $\mathrm{M_{GK}}$ *de guillaume-Kenchaff, Loevinger sont implicatives.*

Définition 7 *(a) Une mesure de qualité des règles* μ *sera dite* symétrique *si pour toute règle d'association* $X \rightarrow Y$*, on a* $\mu(Y \rightarrow X) = \mu(X \rightarrow Y)$*.* μ *sera dite* parfaitement symétrique*, si pour toute règle d'association* $X \rightarrow Y$*, on a* $\mu(\overline{X} \rightarrow \overline{Y}) = \mu(X \rightarrow Y)$ *[AZ03].*

(b) Une mesure de qualité des règles μ sera dite orienée, *s'il existe au moins une règle d'association $X \rightarrow Y$ telle que l'on a $\mu(Y \rightarrow X) \neq \mu(X \rightarrow Y)$.*

Ainsi, à titre d'exemple, toute mesure implicative est orientée. Alors que les mesures symétriques ne sont pas orientées. Ainsi, les mesures symétriques Lift, Piatestsky-Shapiro, supp *ne sont pas orientées.*

2.2.2 Quelques critères d'éligibilité d'une MPQ

Une étude approfondie des critères d'éligibilité des mesures de objectives de la qualité de règles est effectuée dans [Vai06], afin de classifier certaines d'entre elles en empruntant une classification hiérarchique et une analyse factorielle des correspondances. Nous utiliserons dans la section qui suit une classification par le treillis de l'analyse de concepts formelle. Le résultat de ceci pourra être confronté à celui obtenu dans [Vai06]. Avant tout, voici quelques exemples de critères souhaités ainsi évoqués pour apprécier une mesure de qualité de règle d'association :

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

(2) ***Possibilité de choix*** *(Poscho) :*
Une mesure doit impérativement permettre de choisir entre $X \rightarrow Y$ et $X \rightarrow \overline{Y}$ [Fre99].

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

(4) ***Sémantique d'implication****(SemImp) :*
Une mesure doit évaluer de la même façon $X \rightarrow Y$ et $\overline{Y} \rightarrow \overline{X}$, i.e., implicative [Kod99, DRT07].

(5) ***Sensible au contre-exemple****(SbCtr) :*
L'évaluation de l'intérêt d'une règle peut se mesurer favorablement en fonction du nombre élevé d'exemples de la règle ou en fonction de nombre faible de ses contre-exemples [Fre99].

(7) ***Vérification des situations de références****(Vfsit) :*
L'utilisation de mesures de qualité prenant des valeurs positives pour les règles intéressantes permet de se rapprocher des a priori *de l'utilisateur sur la notion de qualité.*
Selon Piatetstky-Shapiro [PS91a], une bonne mesure μ de qualité de la règle $X \rightarrow Y$ doit être :

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

(8) ***Nulle en cas d'indépendance***(*NulInd*) : $\mu(X \rightarrow Y) = 0$ *en cas de l'indépendance entre la prémisse et le conséquent d'une règle, i.e.,* $p(Y'|X') = p(Y')$ *; Ce qui va permettre d'éliminer systématiquement les règles dont prémisse et conséquent sont ind/'ependants ou "proches" de l'ind/'ependance statistique.*

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

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

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

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

(13) ***Fixation facile d'un seuil***(*FixFS*) :
Les mesures de qualité retenues pour extraire et classifier les règles d'association doivent pouvoir être utilisées avec un seuil d'élagage de manière à éliminer toutes les règles qui n'intéressent pas l'utilisateur [LMV$^+$*04]. Les mesures ayant un sens concret pour l'utilisateur, ainsi que les mesures normalisées et ayant un caractère statistique se prête bien à la détermination d'un seuil d'élagage. Ce seuil peut être fixé par l'utilisateur soit avant la phase d'extraction des règles d'association, soit lors d'une phase de post-élagage des règles. Néanmoins, lorsque le seuil est déterminé a priori par l'utilisateur, ce seuil ne prend pas en considération la nature des données et peut conduire à des résultats ne présentant pas toujours les données. L'utilisation de seuils d'élagage calculés directement à partir des données peut permettre d'éviter ce problème : il est donc souhaitable que le seuil soit statistique. De tels seuils peuvent être obtenus à partir des valeurs moyennes observées sur les données. Une méthode classique en fouille de données [Ler84] consiste à centrer et réduire les valeurs observées.*

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

(15) ***Fonction décroissante de la taille de prémisse/conséquent***
(FoDTp) : une bonne mesure de qualité doit être une fonction décroissan- te de la taille de la prémisse (resp. du conséquent) lorsque les autres paramètres

sont fixés [Fre99] .

Eu égard à la littérature sur la fouille des règles d'association, il s'avère que les mesures Support et Confiance sont les plus utilisées par les différentes méthodes de la fouille des règles d'association. Toutefois, depuis ces dernières années, l'utilisation de ces mesures suscite plusieurs critiques. En effet, d'une part, ces mesures peuvent sélectionner certaines règles sans intérêts (cas de l'indépendance entre la prémisse et le conséquent d'une règle si elle a une valeur de Confiance dépasse le seuil minimum Confiance) [LT04, BMUT97], d'autre part, la mesure Support, considérée comme moteur de processus d'extraction, écarte les règles ayant un petit Support alors que certaines peuvent avoir une très forte Confiance et présenter un réel intérêt : les pépites de connaissances [AZ03]. Pour tenter de pallier cet inconvénient, plusieurs mesures de qualité ont été proposées. Ce qui engendre de nouveaux problèmes, entre autres, le choix de mesure(s) utilisée(s) pour l'extraction des règles d'un contexte de la fouille de données. Il se pose ainsi naturellement le problème de choix de la mesure de qualité à utiliser pour sélectionner les règles intéressantes à partir d'un contexte de la fouille de données ? Ce qui explique la suggestion de ces critères de conception ou construction et de guide dans le choix d'une mesure de qualité à utiliser afin de capturer des règles intéressantes. Cependant, comme nous le verrons dans la suite du document, il est très difficile de trouver une mesure vérifiant l'ensemble de ces critères.

2.2.3 Exemples de mesures probabilistes de qualité

Comme ce tableau d'exemples l'indique (Cf. Tableau 2.4), la littérature atteste que sur la façon d'attribution un nom à une mesure de qualité, trois catégories se profilent : il y a les mesures qui portent le nom de l'auteur ou du groupe d'auteurs (telles : mesure de Lovinger, Laplace, Sebag, etc.), puis celles dont le nom reflète les propriétés mathématiques effectives ou la sémantique intrinsèque de l'indice (cest le cas de : indice d'implication (IndImp), mesure de similarité, coéfficient de corrélation linéaire, implication orientée normalisée (ION), etc.), et les mesures dont le nom exprime le souhait (pour ne pas dire le marketing) de son auteur (comme : conviction, confiance, facteur de certitude, etc.*). Cette liste n'est pas exhaustive, car il existe d'autres mesures (voir ([Vai06], [GH06], [HH99]). Nous pensons qu'il est plus objectif de prioriser cette deuxième façon, et que c'est en cas de chevauchement seulement qu'on suffixerait par le nom de l'auteur pour nommer une mesure de qualité de règle.* C'est ainsi par exemple que S. Ferré [Fer06] adopte la dénomination *Confiance de Guillaume* notée $Conf_G$ pour ladite mesure dite de Guillaume-Kenchaff abrégée M_{GK} [Gui00], alors que dans [TR05], nous employons la terminologie d'Implication orientée normalisée, abrégée *ION*, et *conditional probability incremental ratio* ou *CPIR* dans [WZZ04].

Cette diversité d'appelations d'un même concept au sein d'un même domaine témoigne de la jeunesse de cette discipline de data mining. Il y convient donc d'avoir une action d'harmonisation et d'unification devant ce foisonnement des mesures de qualité de règle d'association.

Des travaux allant dans ce sens existent déjà(voir par exemple [Vai06]). Il y est proposé d'adopter une paramétrisation de mesure en cas de non constance à l'équilibre (ou à l'indétermination) et à l'indépendance : ce qui amène au centrage et / ou à la réduction de ladite mesure initiale.

Numéro	Mesure	Expression	Reférence
1	Support	$p(X' \cap Y')$	[AIS93]
2	Confiance	$p(Y'\|X')$	[AIS93]
3	M_{GK}	$\frac{p(Y'\|X')-p(Y')}{1-p(Y')}$ si $p(Y'\|X') \geq p(Y')$ $\frac{p(Y'\|X')-p(Y')}{p(Y')}$ si $p(Y'\|X') \leq p(Y')$	[Gui00]
4	Rappel	$p(X'\|Y')$	[LFZ99]
5	Lift	$\frac{p(Y'\|X')}{p(Y')}$	[BMS97]
6	Laverage	$p(Y'\|X') - p(X')p(Y')$	[GH06]
7	Confiance centrée	$p(Y'\|X') - p(Y')$	[LT04]
8	Facteur de certitude	$\frac{p(Y'\|X')-p(Y')}{1-p(Y')}$	[GH06]
9	Laplace	$\frac{n.p(X'\cap Y')+1}{np(X')+2}$	[Goo65]
10	ϕ-coefficient	$\frac{p(X'\cap Y')-(X')p(Y')}{\sqrt{p(X')p(Y')p(\overline{X}')p(\overline{Y}')}}$	[Ler81]
11	Piatetsky-Shapiro	$p(X' \cap Y') - p(X')p(Y')$	[PS91a]
12	Cosinus	$\frac{p(X'\cap Y')}{\sqrt{p(X')p(Y')}}$	[HGB05a]
13	Accuracy	$P(X' \cap Y') + p(\overline{X}' \cap \overline{Y}')$	[GH06]
14	Moindre Contradiction	$\frac{p(X'\cap Y')-p(\overline{Y}\cap\overline{Y}')}{p(Y')}$	[AK02]
15	Loevinger	$1 - \frac{p(X'\cap\overline{Y}')}{p(X')p(\overline{Y}')}$	[Loe47]
16	Kappa	$2\frac{p(X'\cap Y')-p(X')p(Y')}{p(X')+p(Y')-2p(X')p(Y')}$	[Coh60]
17	Indice d'Implication	$\sqrt{n}\frac{p(X'\cap Y')-p(X')p(Y')}{\sqrt{p(X')p(Y')}}$	[LGR81]
18	Spécificité	$p(\overline{Y}'\|\overline{X}')$	[LFZ99]
19	Fiabilité Negative	$p(\overline{X}'\|\overline{Y}')$	[LFZ99]
20	Zhang	$\frac{p(X'\cap Y')-p(X')p(Y')}{max\{p(X'\cap Y')p(\overline{Y}');p(Y')p(X'\cap\overline{Y}')\}}$	[Zha00]
21	Q-Yule	$\frac{p(X'\cap Y')p(\overline{X'\cap Y'}-p(X'\cap\overline{Y}')p(\overline{X}'\cap Y')}{p(X'\cap Y')p(\overline{X'\cap Y'})+p(\overline{X}'\cap Y')p(X'\cap\overline{Y}')}$	[GH06]
22	Y-Yule	$\frac{\sqrt{p(X'\cap Y')p(\overline{X'\cap Y'})}-\sqrt{p(X'\cap\overline{Y}')p(\overline{X}'\cap Y')}}{\sqrt{p(X'\cap Y')p(\overline{X'\cap Y'})}+\sqrt{p(X'\cap\overline{Y}')p(\overline{X}'\cap Y')}}$	[GH06]
23	J-mesure	$p(X' \cap Y')log(\frac{p(X'\cap Y')}{p(X')p(Y')})$ $+p(X' \cap \overline{Y}')log(\frac{p(X'\cap\overline{Y}')}{p(X')p(\overline{Y}')})$	[GS88]
24	Multiplicateur des côtes	$\frac{p(X'\cap Y')p(Y')}{p(X'\cap\overline{Y}')p(Y')}$	[Lal02]
25	Sebag	$\frac{p(Y'/X')}{p(\overline{Y}/X')}$	[SS88]
26	Conviction	$\frac{p(X').p(\overline{Y}')}{p(X'\cap\overline{Y}')}$	[BMS97]
27	Odd Ratio	$\frac{p(X'\cap Y').p(\overline{X}'\cap\overline{Y}')}{p(\overline{X}'\cap Y')p(X'\cap\overline{Y}')}$	[HGB05a]
28	Klosgen	$\sqrt{p(X' \cap Y')}(p(Y'/X') - p(Y'))$	[HGB05a]
29	Gain Informationel	$log\frac{p(X'\cap Y')}{p(X')p(Y')}$	[CH90]
30	Exemples contre-exemples	$1 - \frac{p(X'\cap\overline{Y}')}{p(X'\cap Y')}$	[GH06]

TAB. 2.4 – Exemples de mesures de qualité

Remarquons que la plupart de ces mesures de qualité de règle peuvent s'exprimer en fonction de la probabilité conditionnelle sachant la prémisse du conséquent(aux extensions respectifs près), dite la confiance de la règle par Agrawal et al. (1993). Alors que la probabilié conditionnelle sachant le conséquent de la prémisse mesure ce qui est appelée la complétude de la règle selon Freitas (1999) ou son rappel selon Azé et Kodratoff (2002)[AK02]. Cest dire l'importance du concept de probabilité conditionnelle pour ces types de mesures d'intérêt de règle d'association.

Par ailleurs, en nous basant sur ces quatorze critères d'elligibilité rappelés cidessus, à l'aide d'un treillis de concepts formels adapté, nous procédons à une classification implicative de ces quelques mesures dans le paragraphe qui suit.

2.2.4 Classification des mesures de qualité selon un treillis

L'analyse du treillis obtenu par la trentaine de mesures eu égard à ces quatorze critères d'éligibilité montre le résultat suivant :

D'une part, on retrouve les deux regroupements "mesures symétriques" versus "mesures orienées" déjà obtenus par les études de Vaillant [Vai06].

D'autre part, et en plus de ce regroupement, il s'avère qu'il existe des implications vers la mesure M_{GK} pour les mesures orientées. Ce qui s'interprête par le fait que, de façon générale, les propriétés vérifiées par les autres mesures orientées sont aussi respectées par M_{GK}. Par conséquent, cette dernière occupe une place importante parmi les mesures orientées et implicatives.

Dans le chapitre suivant nous identifierons, entre autres, une autre classification des MPQs et les MPQs qui se prêtent mieux à produire parallèlement des rêgles d'association positives et négatives.

Chapitre 3

Normalisation d'une mesure probabiliste de qualité

3.1 Mesure probabiliste de qualité normalisée

3.1.1 Motivations

Dans la présente section, dans le double objectif de comparaison et d'une vision unificatrice des différentes mesures de qualité, suite de nos travaux [Tot03], nous proposons une *normalisation*. Cependant notre approche diffère de celle présentée dans [LFZ99]. Elle est en fait guidée par les propriétés des intervalles réels qui sont rappelés ci-dessous, car il est manifeste que ces mesures de qualité prennent leurs valeurs soit dans un intervalle borné, soit dans un intervalle ouvert non borné. Et dans les deux cas, l'intervalle peut contenir ou ne pas admettre des nombres réels négatifs et n'est pas mis en correspondance avec des situations de références plus ou moins intuitives telles l'incompatibilité, l'indépendance statistique, la nature de dépendance (positive ou négative) et l'implication logique.

3.1.1.1 Une propriété topologique de $\mathbb{R}$: intervalles homéomorphes ou difféomorphes

Définition 8 – *Deux intervalles réels sont dits de même nature, s'ils sont simultanément fermés bornés ou non bornés, semi-ouverts bornés ou non bornés, ou ouverts bornés ou non bornés.*

– *Un homéomorphisme (resp. difféomorphisme) de* $\mathbb{R}$ *est une application réelle continue (resp. dérivable à dérivée continue) bijective dont la réciproque est également continue (resp. dérivable à dérivée continue).*

Thèorème 2 *Deux intervalles réels I et J sonts homéomorphes (resp. difféomorphes), si et seulement s'ils sont de même nature.*

Preuve : Notons $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$

(CN) : Si f :$I \longrightarrow J$ est un homéomorphisme, alors des trois choses l'une :

α) Si I est un segment $[a, b]$ avec a, $b \in \mathbb{R}$, alors J ne peut être qu'un segment [c, d] , car l'image d'un segment par une application continue est un segment, toute application continue sur un segment étant bornée et y atteignant ses bornes (Théorème de Bolzano-Weierstrass : De toute suite bornée de $\mathbb{R}$ on peut extraire une sous-suite convergente).

β) Si $I =]a, b[$, avec a, $b \in \overline{\mathbb{R}}$, alors J ne peut être qu'un intervalle ouvert]a, b[, puisque l'image réciproque d'un ouvert par une application continue est un ouvert.

γ) Si $I =]a, b]$ou $I = [a, b[$, avec a, $b \in \overline{\mathbb{R}}$, alors J ne peut être qu'un intervalle de cette nature, sinon on applique α) ou β). ou la fonction tangente hyperbolique $th : \mathcal{R} \longrightarrow]-1, 1[$ composée avec une fonction simple, comme proposées ci-dessous :

- Cas d'intervalles ouverts : on peut prendre $f = g_1 oth$, avec g_1 une affine du type g donnée ci-dessus pour I intervalle borné, $a + Exp(x)$ ou $a - Exp(x)$ selon que l'intervalle I soit de type $]a, +\infty[$ ou $] - \infty, a[$.
- cas d'intervalles semi-ouverts : en considérant la partie de th à valeurs dans l'intervalle [0, 1[on peut prendre la fonction composée $g_2 oth$ ou $g_3 oth$, où g_2 et g_3 sont 2 applications affines de type g définie ci-après sur $[0, 1[$ et à valeurs dans $\mathbb{R}$, ou $a + x$ ou $-x + a$.

(CS) : Si I et J sont de même nature, il suffit d'exhiber un difféomorphi- sme dans chacun des cas.

Lorsque les extrémités a, b, c et d des intervalles I et J sont des réels, on peut choisir des restrictions de l'application affine

$$g : [a, b] \longrightarrow [c, d] \quad x \longmapsto \frac{d - c}{b - a}(x - a) + c, \tag{3.1}$$

et de noter que g est bien un $\mathcal{C}^\infty$-difféomorphisme de [a, b] sur [c, d]. Si au moins l'une des bornes de I est infinie, il suffit d'utiliser la fonction exponentielle
exp : $\mathbb{R} \longrightarrow]0, +\infty[$, $x \longmapsto e^x$.

3.1.1.2 Une diversité de définitions

"Normer", "normaliser", "centrer" une mesure de qualité des règles sont des termes utilisés plus ou moins vaguement par les chercheurs travaillant dans le domaine de la fouille des règles d'association. Pour tenter de corriger les faiblesses de la mesure Confiance, Lallich et Teytaud [LT04] définissaient la mesure Confiance centrée en enlevant de la Confiance la probabilité du conséquent de la règle pour avoir une référence en cas d'indépendance entre la prémisse et le conséquent de la règle. La mesure de Loevinger [Loe47], l'une des plus anciennes mesures de qualité répertoriées dans le domaine de fouille de données, normalise la Confiance centrée. Elle permet de pallier un des défauts de la Confiance. Brin et al. [BMUT97] préconisaient la conviction qui est une mesure implicative normalisée. Cependant, il s'avère que la normalisation évoquée et souhaitée n'est pas explicitée.

Toutefois, comme la Conviction prend la valeur limite $+\infty$ en cas d'implication logique entre la prémisse et le conséquent d'une règle, elle ne permet pas d'indiquer à partir de quelles valeurs de la Conviction "une règle est dite convaincante". En fait, les mesures de qualité ont des comportements hétérogènes face aux critères souhaités pour une bonne mesure de qualité des règles. On peut observer d'importantes variations entre les formules et des grandes différences dans les ensembles des

valeurs prises par une mesure. Pour tenter de clarifier ceci, considérons le contexte de fouille de données présenté dans le Tableau 3.1 formé de cinq attributs et six entités.

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

TAB. 3.1 – Contexte binaire

Le Tableau 3.2 présente les valeurs prises par quelques mesures de qualité pour certaines règles d'association considérées. On constate que les valeurs prises par

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

TAB. 3.2 – Valeurs prises par quelques mesures de qualité

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

TAB. 3.3 – Comportements de quelques mesures de qualité

ces mesures de qualité sont distribuées entre -1 et $+\infty$.

De plus, certaines mesures de qualité prennent des valeurs positives in- dépendamment du fait que la prémisse favorise le conséquent. En fait, nous avons le Tableau 3.3 qui présente les comportements de ces différentes mesures de qualité dans les situations de référence. Prenons par exemple le cas de la Confiance : elle n'a pas une valeur fixe en cas d'indépendance entre la prémisse et le conséquent d'une règle. Ce

qui entraîne la possibilité de sélectionner des règles où la prémisse et le conséquent sont indépendants et même si la prémisse défavorise le conséquent pourvu qu'elles vérifient les conditions de Support et Confiance, alors que les règles de ce type n'ont aucun intérêt pour l'utilisateur. Certaines mesures ne prennent pas de valeurs fixes à l'implication, ce qui engendre la difficulté de définir un seuil minimum. D'où la possibilité d'écarter certaines règles intérêssantes (en cas de l'implication entre la prémisse et le conséquent).

L'objectif de la normalisation est alors de ramener les valeurs d'une mesure de qualité sur l'intervalle $[-1, 1]$ tout en reflétant les situations de référence telles que l'incompatibilité, la dépendance négative, l'indépendance, la dépendance positive et l'implication logique entre la prémisse et le conséqu- ent d'une règle d'association. Ce qui est théoriquement possible, surtout pour les mesures de qualité bornées. Dans une perspective de généralisation des mesures de qualité de régle, arguant qu'une mesure ne satisfait jamais simultanément la condition de constance en cas d'indépendance et en cas d'équilibre [Vai06] propose la définition suivante :

Définition 9 *Une mesure μ est une mesure généralisée de Confiance, si l'on a l'une des 4 formes suivantes : pour tous motifs X et Y, on a :*

(i) $\mu(X \to Y) = f(P(X'), P(Y'), n)(\text{conf}(X \to Y) - \theta)$

(ii) $\mu(X \to Y) = f(P(X'), P(Y'), n)(\frac{\text{conf}(X \to Y)}{\theta})$

(iii) $\mu(X \to Y) = f(P(X'), P(Y'), n)\frac{1-\text{conf}(X \to Y)}{Conf(X \to Y)}$

(iv) $\mu(X \to Y) = f(P(X'), P(Y'), n)\frac{\text{conf}(X \to Y)-\theta}{1-\theta}$

où $\theta = P(Y')$ si la mesure prend une valeur fixe à l'indépendance, et $\theta = 1/2$ si elle prend une valeur fixe à l'équilibre.

Mais cette définition a le défaut d'ignorer entre autres les autres situations de référence intuitives telles l'incompatibilité et l'implication logique.

3.1.2 Notre approche

Avant de donner la définition proposée pour une mesure normalisée, nous adoptons les définitions suivantes en ce qui concerne les situations de référence intuitives en probabilités.

Soient X et Y des motifs d'un contexte de la fouille de données $(\mathcal{O}, \mathcal{A}, \mathcal{R})$. Par souci de cohérence avec le principe de dualité dans l'analyse des concepts formelle, nous les caractérisons selon les propriétés de leurs extensions respectives X' et Y' en tant qu'événéments de $\mathcal{P}(\mathcal{O})$.

Définition 10 *On dit que :*

(i) X et Y sont incompatibles, si leurs extensions sont incompatibles, c-à-d si $P(X' \cap Y') = 0$) ;

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

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

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

(v) X implique logiquement (totalement) Y si $X' \subseteq Y'$,
soit $P(Y'|X') = 1$.

Ainsi, les quantités $P(Y'|X') - P(Y')$ et $P(X'|Y') - P(X')$ mesurent l'écart à l'indépendance des deux motifs X et Y, et donc évaluent le degré de lien orienté entre ces deux motifs. Notons que ces quantités sont préférables à $ln(P(Y'|X')) - ln(P(Y'))$ et $ln(P(X'|Y')) - ln(P(X'))$, ou aux rapports $\frac{P(Y'|X')}{P(Y')}$ et $\frac{P(X'|Y')}{P(X')}$, ces dernières étant plus complexes.

En général ces deux indicateurs de degré de dépendance statistique ne sont pas égaux, malgré la mutualité de l'*attraction* ou de la *répulsion* selon que le lien est positif ou négatif.

Néanmoins les notions de dépendence positive et dépendance négative sont liées comme le montrent les lemmes ci-dessous.

Lemme 3 *Soient X et Y deux motifs.*

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

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

Notons que les deux quantités $P(Y'|X')$ et $P(X'|Y')$ sont fonctions croissantes du nombre d'exemples $|X' \cap Y'|$, les marginales $P(X')$ et $P(Y')$ demeurant constantes. Par ailleurs, la littérature suggère déjà les cinq principes suivants :

- Les trois principes de Piatetsky-Shapiro (1991) [PS91] : une mesure d'intérêt d'une règle d'association doit être nulle en cas d'indépendance statistique des premisse et conséquent, fonction strictement croissante du nombre d'exemples, les autres paramètres étant fixés, et une fonction strictement décroissante du cardinal du dual de sa prémisse ou décroissante du cardinal du dual de son conséquent, les autres paramètres étant maintenus constants.
- Un quatrième principe de Major et Mangano (1993) [MM93] : Une mesure d'intérêt d'une règle d'association doit être une fonction strictement croissante de sa couverture(i.e. le cardinal de l'intersection des deux extensions), une fois que sa confiance est gardée constante supérieure à une valeur minimale préalablement fixée.
- Le cinquième principe de Freitas (1999) [Fre99] qui corrige le caractère symétrique de l'indice de Piatetsky-Shapiro : Une mesure de qualité d'intérêt d'une règle d'association doit être non symétrique.

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

Définition 11 *Soit $X \rightarrow Y$ une règle d'association. Une mesure de qualité μ est dite* normalisée *si elle vérifie les cinq conditions ci-dessous [Tot03], [DRT07] :*

(i) $\mu(X \rightarrow Y) = -1$, si $P(Y'/X') = 0$;
(ii) $-1 < \mu(X \rightarrow Y) < 0$, si $0 \neq P(Y'/X') < P(Y')$(i.e. X et Y sont négativement dépendants (en répulsion partielle) ;
(iii) $\mu(X \rightarrow Y) = 0$, si P(Y'/X')=P(Y')(i.e. X et Y sont indépendants) ;
(iv) $0 < \mu(X \rightarrow Y) < 1$, si $1 \neq P(Y'/X') > P(Y')$, i.e.si X favorise Y ou Xet Y s'attirent partiellement;
(v)$\mu(X \rightarrow Y) = 1$, si $P(Y'/X') = 1$ (ou encore si X implique totalement Y.

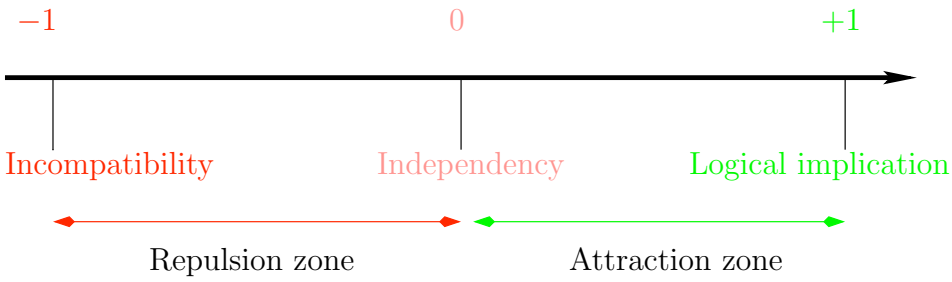

FIG. 3.1 – Distribution of the values of a normalized PQM

Ainsi, il est important de noter que, par essence même, une mesure normalisée possède la sémantique de lien orienté, qui peut s'interpréter en termes de taxonomie, implicite dans un sylogisme tel "Si X, alors Y". C'est un indice de quasi-implication ou de l'implication statistique, pour reprendre la terminologie de l'analyse implicative de Gras. La distribution des valeurs d'une mesure normalisée se représente schématiquement comme indiquée dans la figure Figure 3.1.

Or, avec cette définition, comme nos échanges avec Jean DIATTA l'ont suggéré, une fonction de type $\mu_{(1/3,2/5)}$ telle que :

$$\mu_{(1/3,2/5)}(X \rightarrow Y) = \begin{cases} 1, & si \quad X \quad \text{implique totalement} \quad Y \\ 1/3, & si \quad X \quad \text{implique partièlement} \quad Y \\ 0, & si \quad P(Y'/X') = P(Y') \\ -2/5, & si \quad 0 \neq P(Y'/X') \leq \quad P(Y') \\ -1, & si \quad P(Y'/X') = 0 \end{cases}$$

est une mesure normalisée, quoique discontinue sur les intervalles $[0, 1]$ et $[-1, 0]$. Cet exemple de mesure normalisée se généralise naturellement en $\mu_{(a,r)}$, où $a \in]0, 1[$ et $r \in]-1, 0[$. Par souci de cohérence avec les réalités statistiques de données, nous sommes alors conduits à introduire une définition qui prend en compte la continuité sur l'intervalle $[-1, 1]$.

Notons par $\mathcal{N}$ l'ensemble des mesures probabilistes de qualité normalisées.

Définition 12 *On appelle mesure de qualité normalisée continue toute mesure de qualité normalisée de $\mathcal{N}$ qui est fonction continue du nombre de contre-exemples (ou d'exemples), ou ce qui revient au même, une fonction continue de la probabilité conjointe des extensions de la premisse et du conséquent d'une règle d'association, ou même de la probabilité conditionnelle sachant la prémisse du conséquent, aux marginales fixées.*

Remarque 3 *Notons que cette définition de mesure normalisée est différente de celle proposée dans Hilderman Robert et Hilderman Howard (1999) [HH99] et dans [AZ04], selon laquelle il suffit que les domaines de valeurs de la mesure en question soit l'intervalle $[-1, +1]$ sans prise en compte des situations de références intuitives évoquées ci-dessus. Elle est également différente du concept d' indice statistique normalisé proposé par Lerman et Azé (2003) : un tel indice n'est pas nécessairement une mesure normalisée au sens de notre présente approche.*

Aux trois critères de Piatetsky-Shapiro [PS91a], nous avons ajouté deux conditions supplémentaires, à savoir la valeur -1 *en cas d'incompatibilité et la valeur* 1 *en cas de l'implication logique de la prémisse sur le conséquent de la règle, afin d'encadrer les valeurs prises par une mesure de qualité, ces deux réels étant opposés comme l'évement impossible et l'événement certain. Cet encadrement permet d'indiquer si l'attraction ou la répulsion entre la prémisse et le conséquent de la règle est forte ou faible. Par exemple, pour une règle d'association* $X \rightarrow Y$*, une valeur de mesure de qualité voisine de* 1 *indique que l'attraction est forte entre la prémisse et le conséquent, donc la règle est intéressante. Par contre, une valeur de mesure voisine de* -1 *indique que la répulsion est forte entre la prémisse et le conséquent, dans ce cas les règles négatives à droite et contraposées* $X \rightarrow \overline{Y}$ *et* $Y \rightarrow \overline{X}$ *pourraient être intéressantes. Ce qui n'est pas le cas si la mesure n'est pas bornée.*

3.1.3 Propriétés des mesures probabilistes de qualité normalisées

Dans la suite $\mathcal{C}(\mathcal{N})$ désigne l'ensemble des mesures de qualité normalisées continues sur $[-1, 1]$.
Alors $\mathcal{C}(\mathcal{N}) = \mathcal{N} - \{\mu_{(a,r)}$, tel que $a \in]0, 1[$ et $r \in]-1, 0[\}$.

Une mesure normalisée continue est nécessairement une mesure généralisée de la confiance, le paramètre θ étant ici égal à la probabilité de l'extension du conséquent. Par souci de commodité, sauf mention expresse, nous considérons essentiellement les mesures normalisées continues.

De plus, comme sa définition le sous-entend, avec la contrainte de nullité en cas d'indépendance, avec l'intégrité du corps $\mathbb{R}$, on a la condition nécessaire suivante : une mesure μ normalisée continue doit s'exprimer en deux morceaux de la façon suivante :

Notons $EI(X \rightarrow Y) = (P(Y'/X') - P(Y'))$ l'écart à l'indépendance de Y sur X.

Proposition 4 *Si une mesure de qualité* μ *est normalisée, alors :*

$$\mu(X \rightarrow Y) = \begin{cases} f(n, P(X'), P(Y'), P(X' \cap Y'))EI(X \rightarrow Y), & si \quad X \quad fav. \quad Y \\ g(n, P(X'), P(Y'), P(X' \cap Y'))EI(X \rightarrow Y), & si \quad X \quad défav. \quad Y \end{cases}$$

f et g étant deux fonctions réelles strictement positives et inférieures ou égales à 1.

Corollaire 1 *Toute mesure de qualité normalisée continue produit des règles plus pertinentes que la mesure confiance, cette dernière pouvant générer des règles à prémisse et conséquent indépendants.*

Maintenant, observons ce qui se passe aux deux autres situations de référence, le cas limite d'implication logique où μ doit prendre la valeur $+1$, et son antipode, le cas d'incompatibilité où la mesure normalisée doit valoire -1.

Alors, par la première contrainte, qui correspond au cas où $X' \subset Y'$, donc X favorise Y, nécessairement :

$\mu^f(X{\rightarrow}Y) = f(n, P(X'), P('Y'))(1 - P(Y')) = 1$ en cas limite de règle logique, donc en cas d'attraction mutuelle il existe un facteur positif λ tel que

$f(n, P(X'), P'Y')) = \lambda(n, P(X'), P('Y'), P(X' \cap Y')) \times \frac{1}{(1-P(Y'))}$,
avec comme contrainte : en cas d'implication logique λ prend la valeur $+1$.
En raisonnant de façon analogue, la contrainte du cas d'incompatibilité donne que : $g(n, P(X'), P(Y')) = \frac{-1}{P(Y')}$ en cas limite d'incompatibilité, et sur la zone de répulsion, il existe aussi un facteur multiplicatif positif β tel que $g(n, P(X'), P(Y'), P(X' \cap Y')) = -\beta(n, P(X'), P(Y'), P(X' \cap Y')) \times \frac{1}{P(Y')}$.
Donc, par continuité de chaque morceau de ladite mesure normalisée, μ^f (composante favorable) et μ^d (partie défavorable), nous aboutissons à la mesure μ définie explicitement par : pour toute règle d'association $X{\rightarrow}Y$,

$$\mu(X{\rightarrow}Y) = \begin{cases} \lambda(n, P(X'), P('Y'), P(X' \cap Y')) \times \frac{P(Y'/X') - P(Y')}{1-P(Y')}, & si \quad X \quad fav. \quad Y \\ \beta(n, P(X'), P(Y'), P(X' \cap Y')) \times \frac{P(Y'/X') - P(Y')}{P(Y')}, & si \quad X \quad \text{défav.} \quad Y, \end{cases}$$

λ et β étant deux fonctions réelles à valeurs dans $]0, 1]$.
Finalement, on obtient la proposition 5 de la décomposition canonique d'une mesure normalisée continue suivante.

Proposition 5 *Toute mesure probabiliste de qualité normalisée μ peut se décomposer canoniquement en fonction de* $\mathrm{M_{GK}}$ *de la façon suivante :*
$\mu = \lambda \times {\mathrm{M_{GK}}}^f \times \mathbf{1}_f + \beta \times {\mathrm{M_{GK}}}^d \times \mathbf{1}_d$,
*où $\mathbf{1}_f$ désigne l'indicatrice de l'événement "*Prémisse favorise Conséquent*", $\mathbf{1}_d$ l'indicatrice de l'événement "*Prémisse défavorise Conséquent*", λ et β étant deux fonctions réelles à valeurs dans* $]0, 1]$.

Nous appelons ceci la $\mathrm{M_{GK}}$-*décomposabilité canonique* d'une mesure probabiliste de la qualité normalisée μ. Soit : $\mu^f = \lambda \times {\mathrm{M_{GK}}}^f$ et $\mu^d = \beta \times {\mathrm{M_{GK}}}^f$.
Alors une mesure normalisée continue μ se décompose ainsi :
$\mu = \mu^f \mathbf{1}_f + \mu^d \mathbf{1}_d$.
Cependant, la réciproque est fausse.
En effet, comme nous le verrons dans les paragraphes 3.2.1 et 3.3.3, par exemple, la mesure continue $\phi - coefficient$ n'est pas normalisée, mais se décompose selon $\mathrm{M_{GK}}$:
$\phi(X{\rightarrow}Y) = \sqrt{\frac{P(X')P(\overline{Y'})}{P(X')P(Y')}} {\mathrm{M_{GK}}}^f(X{\rightarrow}Y)\mathbf{1_f}(X{\rightarrow}Y) + \sqrt{\frac{P(X')P(Y')}{P(X')P(Y')}} {\mathrm{M_{GK}}}^d(X{\rightarrow}Y)\mathbf{1_d}(X{\rightarrow}Y)$.

Cette décomposition canonique en "partie favorable - partie défavorable" autorise à considérer une mesure normalisée μ comme une fonction vectorielle à deux dimensions, soit $\mu = (\mu^f, \mu^d)$. Ce qui devrait conduire à une interprétation géométrique d'une mesure probabiliste de la qualité normalisée continue. Or, il est facile de vérifier que $\mathrm{M_{GK}}$ est bien normalisée ; donc en fait $\mathrm{M_{GK}}$ correspond au cas simple et maximal où les coéfficients sont égaux à 1 : soit $\lambda = \beta = constante = 1$.

Ainsi, retenons que, malgré les cinq contraintes de normalisation, ce résultat fournit un moyen pour construire une mesure normalisée, aux facteurs λ et β près. Ainsi, les mesures normalisées continues peuvent toutes s'exprimer en une transformée de $\mathrm{M_{GK}}$ et donc de la mesure pionnière $Confiance$.

Exemple 1 – *Il est facile de voir que les deux premières mesures de qualité suivantes sont des mesures normalisées continues.*
– *La mesure de qualité* $\mathrm{M_{GK}}$ *[Gui00] définie par :*

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

– *La mesure de Zhang [Zha00] définie par :*

$$Zhang(X{\rightarrow}Y) = \frac{P(X' \cap Y') - P(X')P(Y')}{max\{P(X' \cap Y')P(\overline{Y}'); P(Y')P(X' \cap \overline{Y}')\}} [Zha00].$$

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

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

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

Par ailleurs, concernant les coéfficients, notons que la mesure de Zang définie ci-dessus correspond au cas où :
$\lambda(n, P(X'), P('Y'), P(X' \cap Y')) = \frac{1}{P(Y'/X')}$
et $\beta(n, P(X'), P(Y'), P(X' \cap Y')) = \frac{1}{1-P(Y'/X')}$.
Au final, à titre indicatif, la mesure normaliée de Zang s'explicite et s'interpête plus facilement sous cette forme éclatée :

$$Zang(X{\rightarrow}Y) = \begin{cases} \frac{\mathrm{M_{GK}}(X{\rightarrow}Y))}{P(Y'/X')}, & si \quad X \quad fav. \quad Y \\ \frac{\mathrm{M_{GK}}(X{\rightarrow}Y))}{1-P(Y'/X')}, & si \quad X \quad \text{défav.} \quad Y \end{cases}$$

Soit :
$Zang(X{\rightarrow}Y) = \frac{\mathrm{M_{GK}}^f(X{\rightarrow}Y)}{P(Y'/X')}\mathbf{1}_f(X{\rightarrow}Y) + \frac{\mathrm{M_{GK}}^d(X{\rightarrow}Y)}{1-P(Y'/X')}\mathbf{1}_d(X{\rightarrow}Y)$.

Par conséquent, contrairement à ce qu'en pensent certains auteurs, les deux mesures de qualité $Zang$ et $\mathrm{M_{GK}}$ sont bien différentes.

3.1.4 Opérations sur les mesures probabilistes de qualité normalisées

Introduisons les lois de composition suivantes dans $\mathcal{N}$.

Définition 13 – **Addition :** $\forall \mu,\ \nu \in \mathcal{N},\ \mu \oplus \nu = \frac{(\mu^f+\nu^f)}{2} + \frac{(\mu^d+\nu^d)}{2}$
– **Addition barycentrique ou combinaison linéaire convexe :**
$\forall \mu, \nu \in \mathcal{N},\ \forall a, b \in \mathbb{R}_+^*,\ a\mu \oplus_B b\nu = \frac{(a\mu^f+b\nu^f)}{a+b} + \frac{(a\mu^d+b\nu^d)}{a+b}$.
– **Produit :**
$\mu\mu' = \mathbf{1}_f\mu^f\mu'^f - \mathbf{1}_d\mu^d\mu'^d$

- **Puisance : pour** $\alpha, \beta > 1$**,**
  $\mu^{\alpha} = \mathbf{1}_f(\mu^f)^{\alpha} + \mathbf{1}_d(-1)^{\alpha-1}(\mu^d)^{\alpha}$
  $\mu^{(\alpha,\beta)} = \mathbf{1}_f(\mu^f)^{\alpha} + \mathbf{1}_d(-1)^{\gamma}(\mu^d)^{\beta}$**, avec** $\gamma = 1$ **si** β **est pair et 0 sinon.**
- **Sup et Inf :** $\mu \vee \mu' = \mathbf{1}_f\mu^f \vee \mu'^f + \mathbf{1}_d\mu^d \vee \mu'^d$, $\mu \wedge \mu' = \mathbf{1}_f\mu^f \wedge \mu'^f + \mathbf{1}_d\mu^d \wedge \mu'^d$, $\mu \ltimes \mu' = \mathbf{1}_f\mu^f \vee \mu'^f + \mathbf{1}_d\mu^d \wedge \mu'^d$, $\mu \rtimes \mu' = \mathbf{1}_f\mu^f \wedge \mu'^f + \mathbf{1}_d\mu^d \vee \mu'^d$

Remarquons d'abord que l'addition $\oplus$ est bien un cas particulier de l'addition barycentrique $\oplus_B$; ce qui rassure sur la cohérence de ces deux lois de composition.

Proposition 6 *(i)* $\mathcal{N}$ *est stable par combinaison linéaire convexe* $\oplus_B$ *(ou addition barycentrique), par le produit, par élévation à une puissance entière, par* $\vee$ *et* $\wedge$*, ces opérations étant commutatives. C'est donc un monoïde.*
De plus, on a : $|\mu^{\alpha}| < |\mu^{\alpha-1}|$ *et* $|\mu\mu'| < |\mu \wedge \mu'|$
(ii) $\mathcal{N}$ *est stable par enveloppe supérieure et par enveloppe inférieure, sous la forme :*
$\forall \mu, \nu \in \mathcal{N}$, $max(\mu, \nu) = max(\mu^f, \nu^f)\mathbf{1}_f + max(\mu^d, \nu^d)\mathbf{1}_d \in \mathcal{N}$
et $min(\mu, \nu) = min(\mu^f, \nu^f)\mathbf{1}_f + min(\mu^d, \nu^d)\mathbf{1}_d \in \mathcal{N}$
(iii) $\forall \mu \in \mathcal{C}(\mathcal{N}), \forall m \in \mathbb{N}, \forall n \in \mathbb{N}^*, (\mu^f)^n\mathbf{1}_f + (\mu^d)^{2m+1}\mathbf{1}_d \in \mathcal{C}(\mathcal{N})$,
$\mu^n \in \mathcal{C}(\mathcal{N}), \mu^{(n,m)} \in \mathcal{C}(\mathcal{N})$.
(iv) $\mathcal{C}(\mathcal{N})$ *est stable par les opérations addition* $\oplus_B$*, produit,* $\vee$ *et* $\wedge$*.*

Démonstration : Vérifions la stabilité de l'addition barycentrique dans (i), les autres étant immédiates. Soient X et Y deux motifs.

$(\mu^f + \nu^f)(X \to Y)/2 = (\mu^f(X \to Y) + \nu^f(X \to Y))/2 = 1$, si X implique logiquement Y, et zéro en cas d'indépendance.

Or $0 < (\mu^f(X \to Y), \nu^f(X \to Y)) < 1$, en cas d'attraction mutuelle,
donc $0 < \mu^f(X \to Y) + \nu^f(X \to Y) < 2$.
De même pour l'autre composante, $-2 < \mu^f(X \to Y) + \nu^f(X \to Y) < 0$.
Pour l'addition barycentrique, il suffit d'examier le cas favorable : des deux doubles inégalités $0 < (\frac{a\mu^f(X \to Y)}{a+b}, b\frac{\nu^f(X \to Y)}{a+b}) < (\frac{a}{a+b}, \frac{b}{a+b})$
résulte $0 < \frac{a\mu^f(X \to Y)}{a+b} + b\frac{\nu^f(X \to Y)}{a+b} < \frac{a+b}{a+b} = 1$. Ce qui démontre la stabilité de $\mathcal{N}$ pour l'addition $\oplus$ et pour l'addition normalisée $\oplus_B$.
Les autres propriétés sont immédiates □
De cette proposition 6 affirmant la stabilité de $\mathcal{C}(\mathcal{N})$ pour cette dizaine d'opérations algébriques résulte naturellement la proposition 7 ci-dessous.

Proposition 7 *Il existe une infinité de mesures de qualité normalisées continues au sens de la définition de la présente approche.*

Remarque 4 *On peut vérifier que l'opération* combinaison linéaire convexe $\oplus_B$ *ne possède malheureusement pas une propriété régularisante d'une mesure de qualité normalisée discontinue. Cependant la stabilité de* $\mathcal{C}(\mathcal{N})$ *par combinaison linéaire convexe fournit déjà un moyen intéressant pour construire une mesure normalisée continue, par exemple satisfaisant une propriété supplémentaire souhaitée. La propriété (iii) de cette proposition 6 fournit un autre moyen pour construire une mesure normalisée continue une fois que l'on en dispose une. Notons ici que le rapport* $\frac{(\mu^f)^{n+1}}{(\mu^f)^n} = \mu^f$ *étant positif et strictement inférieur à* 1*; ce qui fournit ainsi un autre moyen d'obtenir une mesure de qualité normalisée plus sélective : il suffit d'augmenter la puissance de la composante favorable* μ^f*, en faisant attention au seuil préalablement fixé. Cependant il s'y pose un problème d'optimisation sur le "bon choix" de la puissance et du seuil, pour ne pas rater les règles intéressantes.*

Corollaire 2 *Pour deux mesures normalisées* μ *et* ν*, les composantes de leur somme sont telles que :* $\forall a, b \in \mathbb{R}_+^*$*, on a :*
$(a\mu \oplus_B b\nu)^f = \frac{(a\mu^f + b\nu^f)}{a+b} = (\frac{(a\lambda_\mu + b\lambda_\nu)}{a+b})\mathrm{M_{GK}}^f$
et $(a\mu \oplus_B b\nu)^d = \frac{(a\mu^d + b\nu^d)}{a+b} = (\frac{(a\beta_\mu + b\beta_\nu)}{a+b})\mathrm{M_{GK}}^d$.

Reprenons les exemples de mesures normalisées continues *Zang* et $\mathrm{M_{GK}}$. Leur "somme" est définie par :
$(Zang \oplus_B \mathrm{M_{GK}})(X \to Y) = \frac{1}{2}(1 + \frac{1}{P(Y'/X')})(\mathrm{M_{GK}}^f \mathbf{1}_f)(X \to Y) + \frac{1}{2}(1 + \frac{1}{1-P(Y'/X')})(\mathrm{M_{GK}}^d \mathbf{1}_d)(X \to Y)$.
Tout ceci montre que la mesure $\mathrm{M_{GK}}$ occupe le rôle de "base" ou de "noyau" au sein de l'ensemble $\mathcal{C}(\mathcal{N})$ des mesures normalisées continues : elle est vraisemblablement la mesure normalisée la "plus simple".

Puisque $n_{XY} = n_X - n_{X\overline{Y}}$, pour tous motifs X et Y, par rapport à la notion de contre-exemple qui signifie ici contradicteur d'implication, la proposition suivante est immédiate.

Proposition 8 *La mesure de la qualité de règles normalisée* $\mathrm{M_{GK}}$ *est une fonction strictement décroissante du nombre de contre-exemples sur le conséquent d'une règle présents dans sa prémisse.*

Nous donnons d'autres propriétés mathématiques de $\mathrm{M_{GK}}$ après avoir identifié la condition nécessaire et suffisante pour qu'une mesure normalisée lui soit associée par homéomorphie affine.
Avant de caractériser une mesure normalisable, voyons une autre propriété d'une mesure normalisée continue. Il résulte du lemme (3) ci-dessus que :

Proposition 9 *Si une mesure de qualité* μ *est normalisée continue, alors* μ *manipule systématiquement les règles négatives au même titre que les règles positives.*

En effet, dans le cas où un motif X défavorise un motif Y, alors X favorisant $\overline{Y}$, et $\overline{X}$ favorisant Y, la dépendance entre X et $\overline{Y}$ demeure alors positive, de même pour le lien entre $\overline{X}$ et Y.
Par conséquent, comme l'on s'intéresse généralement aux liens positifs entre deux motifs, théoriquement du moins, la mesure normalisée M_{GK} apparaît bien adaptée pour extraire pertinemment les règles d'associations, les règles dont prémisse et conséquent sont indépendants ou proches de l'indépendance étant systématiquement évitées.

Cependant, il est évident que ce ne sont pas toutes les mesures disponibles dans la littérature qui sont normalisées. C'est le cas par exemples de Confiance et Support. Par conséquent, se pose la question de savoir l'existence de moyen pour normaliser une mesure de qualité quelconque.

3.2 Processus et caractérisation de la normalisation

Nous nous restreignons ici à l'ensemble $\mathcal{C}(\mathcal{N})$ des mesures de qualité normalisées continues. Par ailleurs, comme il existe une bijection affine (donc une une fonction très simple) qui relie deux intervalles bornés de $\mathbb{R}$, notre première démarche de recherche de la normalisée d'une mesure va s'intéresser à la recherche d'une transformation affine ou affine par morceaux ou affine avec coéfficients dynamiques.

Cherchons une condition nécessaire et suffisante pour qu'une mesure de qualité soit normalisable.
Notation et terminologie : Sous réserve de son existence, la mesure de qualité probabiliste normalisée μ_n élément de $\mathcal{C}(\mathcal{N})$, déduite d'une mesure de qualité $\mu \in \mathcal{C}(\mathcal{N})$ est dite la mesure normalisée associée à μ.

Considérons maintenant une mesure de qualité μ continue. Cherchons sa mesure normalisée continue $\mu_n \in \mathcal{C}(\mathcal{N})$ associée à μ, si elle existe.

Pour faciliter l'interprétation d'une règle, la normalisation de μ consisterait à ramener ses valeurs sur l'intervalle $[-1, 1]$ de telle sorte que la valeur -1 corresponde à l'incompatibilité, les valeurs strictement comprises entre -1 et 0 correspondent à la répulsion ou la dépendance négative, la valeur 0 corresponde à l'indépendance, les valeurs strictement comprises entre 0 et 1 correspondent à l'attraction ou à la dépendance positive orientée et la valeur 1 corresponde à l'implication logique entre la prémisse et le conséquent d'une règle $X \rightarrow Y$. Soit x_f (resp. y_f) le coéfficient de multiplication (resp. de centrage) de μ, dans le cas où X favorise Y. De façon similaire, posons x_d (resp. y_d) le coéfficient de multiplication (resp. de centrage) dans le cas où X défavorise Y. On a donc :

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

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

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

Au cas où X défavorise Y, on obtient :

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

Nous pouvons écrire le système d'équations linéaires suivant.

$$\begin{cases} x_f.\mu_{imp}(X{\rightarrow}Y) + y_f = 1 \\ x_f.\mu_{ind}(X{\rightarrow}Y) + y_f = 0 \\ x_d.\mu_{ind}(X{\rightarrow}Y) + y_d = 0 \\ x_d.\mu_{inc}(X{\rightarrow}Y) + y_d = -1 \end{cases} \tag{3.2}$$

L'écriture matricielle de l'équation (3.2) est donnée par l'équation (3.3) :

$$\begin{pmatrix} \mu_{imp}(X{\rightarrow}Y) & 1 & 0 & 0 \\ \mu_{ind}(X{\rightarrow}Y) & 1 & 0 & 0 \\ 0 & 0 & \mu_{ind}(X{\rightarrow}Y) & 1 \\ 0 & 0 & \mu_{inc}(X{\rightarrow}Y) & 1 \end{pmatrix} \begin{pmatrix} x_f \\ y_f \\ x_d \\ y_d \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \end{pmatrix} \tag{3.3}$$

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

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

Pour que l'équation (3.3) admette une solution unique, il faut et il suffit que le déterminant de la matrice M soit fini et non nul. Ainsi, nous avons la caractérisation de mesures de qualité normalisables résultant de l'existence de solution de l'équation (3.3).

Thèorème 3 *Une mesure de qualité μ est normalisable si et seulement si, pour toute règle $X \rightarrow Y$, les conditions suivantes sont vérifées :*
(i) les quantités $\mu_{imp}(X \rightarrow Y)$, $\mu_{ind}(X \rightarrow Y)$ et $\mu_{inc}(X \rightarrow Y)$ sont finies ;
(ii) les inégalités suivantes sont vérifiées $\mu_{imp}(X \rightarrow Y) \neq \mu_{ind}(X \rightarrow Y)$, $\mu_{ind}(X \rightarrow Y) \neq \mu_{inc}(X \rightarrow Y)$.

Démonstration : Le déterminant de la matrice M associée à l'équation (3.3) est égal à $(\mu_{imp}(X \rightarrow Y) - \mu_{ind}(X \rightarrow Y))(\mu_{ind}(X \rightarrow Y) - \mu_{inc}(X \rightarrow Y))$. D'où le théorème énoncé. □

Remarque 5
- *Le système d'équations linéaires (3.2) ne peut pas avoir une infinité de solutions. En effet, si $det(M) = 0$, pour assurer l'infinité de solution, il faut que le second membre du système soit nul. Ce qui n'est pas le cas.*
- *Les règles d'association considérées sont des règles $X \rightarrow Y$ telles que $p(X') \neq 0$, $p(Y') \neq 0$, $p(X') \neq 1$ et $p(Y') \neq 1$. En effet, si $p(X') = 1$ donc les attributs qui constituent X sont présents dans toutes les entités, donc le motif X ne porte aucune information nouvelle à l'utilisateur. Par ailleurs, si $p(X') = 0$ la présence simultanée des attributs qui composent X ne se réalise dans aucune entité, donc le motif X ne porte aucune information nouvelle à l'utilisateur.*

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

Proposition 10 *Soient μ une mesure de qualité normalisable et $X \rightarrow Y$ une règle d'association. Les coéfficients de multiplication et de centrage sont donnés par les expressions ci-dessous :*

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

Remarque 6
1. *Il est à noter que les coéfficients x_f, x_d, y_f et y_d ne dépendent que des probabilités $p(X')$ et $p(Y')$ de la même manière que les quantités $\mu_{imp}(X \rightarrow Y), \mu_{ind}(X \rightarrow Y)$ et $\mu_{inc}(X \rightarrow Y)$.*
2. *De plus, il est facile de voir que $\mathrm{M_{GK}}_n = \mathrm{M_{GK}}$. Ce qui est cohérent avec les résulats ci-dessus. Plus généralement, pour toute mesure normalisée $\mu \in \mathcal{C}(\mathcal{N})$, on a : $\mu_n = \mu$.*
3. *Enfin, observons aussi que pour une mesure $\mu \in \mathcal{C}(\mathcal{N})$ à normalisée $\mu_n = \mathrm{M_{GK}}$, on a la relation explicite inverse :*

$$\mu(X \rightarrow Y) = \begin{cases} \frac{\mathrm{M_{GK}}(X \rightarrow Y) - y_f}{x_f}, & si \quad X \quad favorise \quad Y \\ .\frac{\mathrm{M_{GK}}(X \rightarrow Y) - y_d}{x_d}, & si \quad X \quad défavorise \quad Y \end{cases}$$

Cette relation réciproque va permettre, via $\mathrm{M_{GK}}$, de comparer deux mesures μ et μ'vis-à-vis d'une règle d'association candidate R_1 : par exemple, si

R_1 est valide selon $\mathrm{M_{GK}}$ *et non valides selon μ et μ' alors ces deux dernières sous-évaluent les règles approximatives ; sinon, si R_1 est μ -valide et μ'-non valide, alors seule μ' sous-évalue les règles ; si R_1 est non valide selon* $\mathrm{M_{GK}}$*, mais valide selon μ et μ' alors ces deux dernières sur-évaluent les règles approximatives, etc. Ainsi, la mesure* $\mathrm{M_{GK}}$ *joue un rôle de jauge ou d'arbitre pour deux mesures de qualités qui lui sont associées. Par conséquent, au sein du groupe des mesures de qualités ainsi associées à* $\mathrm{M_{GK}}$ *par normalisation par homéomorphie affine, cette dernière permet une vue comparative et unificatrice.*

3.2.1 Exemples de normalisation de mesures de qualité

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

1. **Support** : $\mathrm{supp}(X \to Y) = P(X' \cap Y')$
   $\mathrm{supp}_{inc}(X \to Y) = 0 \neq -1, \mathrm{supp}_{ind}(X \to Y) = P(X')P(Y') \neq 0, \mathrm{supp}_{imp}(X \to Y) = P(X') \neq 1$, donc $det(M) = P^2(X')P(\overline{X})P(Y')P(\overline{Y}') \neq 0$. D'après le Théorème 3, la mesure non normalisée *Support* est normalisable.

$$\begin{array}{ll} x_f = \frac{1}{P(X')(1-P(Y'))}, & y_f = -\frac{P(X')P(Y')}{P(X')(1-P(Y'))} \\ x_d = \frac{1}{P(X')P(Y')}, & y_d = -1 \end{array}$$

   soit

$$\mathrm{supp}_n(X \to Y) = \begin{cases} \frac{P(X' \cap Y') - P(X')P(Y')}{P(X')(1-P(Y'))} \text{ si } X \text{ favorise } Y \\ \frac{P(X' \cap Y') - P(X')P(Y')}{P(X')P(Y')} \text{ si } X \text{ défavorise } Y \end{cases}$$

   Finalement, on trouve que $\mathrm{supp}_n(X \to Y) = \mathrm{M_{GK}}(X \to Y)$.

2. **Confiance** : $\mathrm{conf}(X \to Y) = P(Y'|X')$
   $\mathrm{conf}_{inc}(X \to Y) = 0 \neq -1, \mathrm{conf}_{ind}(X \to Y) = P(Y') \neq 0$ et $\mathrm{conf}_{imp}(X \to Y) = 1$, donc $det(M) = 1 - P(Y') \neq 0$.
   D'après le Théorème 3, la mesure non normalisée $Confiance$ est normalisable.

$$\begin{array}{ll} x_f = \frac{1}{1-P(Y')}, & y_f = -\frac{P(Y')}{1-P(Y')} \\ x_d = \frac{1}{p(Y')}, & y_d = -1 \end{array}$$

   soit

$$\mathrm{conf}_n(X \to Y) = \begin{cases} \frac{P(X' \cap Y') - P(X')P(Y')}{P(X')(1-P(Y'))} \text{ si } X \text{ favorise } Y \\ \frac{P(X' \cap Y') - P(X')P(Y')}{P(X')P(Y')} \text{ si } X \text{ défavorise } Y \end{cases}$$

   Ainsi $\mathrm{conf}_n(X \to Y) = \mathrm{M_{GK}}(X \to Y)$.

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

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

   soit

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

   D'où $Lift_n(X \to Y) = \mathrm{M_{GK}}(X \to Y)$.
4. **Laplace** : $Lap(X \to Y) = \frac{nP(X' \cap Y')+1}{nP(X')+2}$
   $Lap_{inc} = \frac{1}{nP(X')+2} \neq -1$, $Lap_{ind} = \frac{nP(X')p(Y')+1}{nP(X')+2}$, $Lap_{imp} = \frac{nP(X')+1}{nP(X')+2} \neq 1$,
   donc
   $det(M) = \frac{n^2P^2(X')P(Y')(1-P(Y'))}{(nP(X')+2)^2} \neq 0$.
   La mesure de qualité non normalisée Lap est donc normalisable.

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

   soit

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

   Par conséquent $Lap_n(X \to Y) = \mathrm{M_{GK}}(X \to Y)$.
5. ϕ-**coefficient** : $\phi(X \to Y) = \frac{p(X' \cap Y') - p(X')p(Y')}{\sqrt{p(X')p(Y')p(\overline{X}')p(\overline{Y}')}}$
   $\phi_{inc} = -\sqrt{\frac{p(X')p(Y')}{p(\overline{X}')p(\overline{Y}')}} \neq -1$, $\phi_{ind} = 0$, $\phi_{imp} = \sqrt{\frac{p(X')p(Y')}{p(\overline{X}')p(\overline{Y}')}} \neq 1$, donc
   $det(M) = -\frac{p(X')p(Y')}{(p(\overline{X}')p(\overline{Y}')} \neq 0$.
   La mesure de qualité non normalisée ϕ est donc normalisable.

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

   soit

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

D'où $\phi_n(X \to Y) = \mathrm{M_{GK}}(X \to Y)$.

5. **Jaccard** : $Jac(X \to Y) = \frac{p(X' \cap Y')}{p(X')+p(Y')-p(X' \cap Y')}$
   $Jac_{inc} = 0 \neq -1$, $Jac_{ind} = \frac{p(X')p(Y')}{p(X')p(\overline{Y}')+p(Y')} \neq 0$, $Jac_{imp} = \frac{p(X')}{p(Y')} \neq 1$, donc $det(M) \neq 0$.
   La mesure de qualité non normalisée Jac est donc normalisable.

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

soit

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

On obtient ainsi que $Jac_n(X \to Y) \neq \mathrm{M_{GK}}(X \to Y)$.

D'une manière plus globale, à titre indicatif, on démontre que [Fen07] :

Proposition 11 *(i) Les vingt mesures de qualité* $\mathrm{M_{GK}}$*, Support, Confiance, Rappel, Lift, laverage, Confiance-centrée, Facteur de certitude, Laplace, ϕ-coefficient, Piatetsky-Shapiro, Cosinus, Accuracy, Moindre contradiction, Lovinger, Kappa, Indice d'implication, Spécificité et Fiabilité négative ont leurs normalisées associées égales à* $\mathrm{M_{GK}}$ *;*

(ii) Les cinq mesures de qualité Jaccard, Zhang, Q-Yule, Y-Yule, J-mesure sont normalisées et différentes de $\mathrm{M_{GK}}$ *;*

(iii) Les sept mesures Multiplicateur de côte, Sebag, Conviction, Odd Ratio, Klosgen, Gain informationnel et taux d'exemples contre-exemple ne sont pas $\mathrm{M_{GK}}$*-normalisables par une homéomorphie affine.*

Ce résultat permet de classifier les mesures de qualité de règles d'association au moins en deux groupes dont celui des mesures associées à $\mathrm{M_{GK}}$ sont majoritaires. Il s'avère ainsi possible de comparer les mesures de qualité $\mathrm{M_{GK}}$-normalisables. Le problème reste ouvert quant à la transformation qui permettrait la normalisation des douze autres mesures dans un sens encore à préciser. Vu le rôle que puisse jouer la mesure $\mathrm{M_{GK}}$, nous donnons dans la section suivante d'autres propriétés de cette mesure normalisée.

3.3 Étude particulière de la mesure normalisée $\mathrm{M_{GK}}$

Remarque 7 Quelques travaux existants sur $\mathrm{M_{GK}}$.
Remarquons tout d'abord que cette mesure a été proposée indépendamment dans [Gui00] en 2000 par inspiration à la mesure de Lovinger et dans [WZZ04]

en 2004. Signalons au passage que, grâce à ses propriétés mathématiques, cette mesure a déjà réçu trois dénominations différentes selon les auteurs, à savoir ION par [TRD04, TRD05] démontrant sa propriété d'implication orientée normalisée, CPIR(conditional probability increment ratio) par Wu et al (2004) [WZZ04] par sa formule qui exprime un taux d'incrément de probabilié conditionnelle et par son efficacité dans l'extraction des règles d'association non redondantes, enfin $Conf_G$ (confiance de Guillaume) par S. Ferré (cf. pages 139-140 de [FER02] montrant qu'elle est moins ambigüe et plus intelligible donc plus appropriée que la traditionnelle Confiance d'Agrawal et al.(qui ne distingue pas attraction et répulsion) dans la recherche d'implications approximatives pour constituer un système d'information logique. À notre connaissance, très peu de travaux parlent de cet indice.

3.3.1 Une autre construction de $\mathrm{M_{GK}}$

[TR05], [DRT07]. Selon la modélisation probabiliste adoptée dans notre approche, on peut facilement retrouver cette mesure par une démarche plus directe. En effet, soit deux motifs X et Y du contexte considéré. Deux cas se distinguent selon la nature de lien reliant ces deux motifs, excluant le cas trivial d'indépendance statistique.

Remarque 8 *On obtient facilement les doubles inégalités suivantes.*
(i) Si X favorise Y, then $0 < P(Y'|X') - P(Y') \leq 1 - P(Y')$.
(ii) Si X défavorise Y, alors $-P(Y') \leq P(Y'|X') - P(Y') < 0$.
(iii) "X défavorise Y" équivaut à "X favorise $\overline{Y}$" ; donc $1 - P(Y') < 1 - P(Y'|X')$ si et seulement si $P(\overline{Y'}) < P(\overline{Y'}|X')$.

Il en résulte que d'une part :

$$-1 \leq \frac{P(Y'|X') - P(Y')}{p(Y')} \leq 0, \text{ si } X \text{ défavorise } Y \tag{3.4}$$

et d'autre part :

$$0 \leq \frac{P(Y'|X') - P(Y')}{1 - P(Y')} \leq 1, \text{ si } X \text{ favorise } Y \tag{3.5}$$

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

Définition 14 *Soit X et Y deux motifs d'un contexte de fouille de données. On définit la mesure* $\mathrm{M_{GK}}$ *par*

$$\mathrm{M_{GK}}(X \rightarrow Y) = \begin{cases} \frac{P(Y'|X') - P(Y')}{1 - P(Y')}, & \textit{si } X \textit{ favorise } Y \\ \frac{P(Y'|X') - P(Y')}{p(Y')}, & \textit{si } X \textit{ défavorise } Y. \end{cases} \tag{3.6}$$

Tout d'abord, notons que pour deux motifs X et Y non indépendants, des deux choses l'une : soit il y a attraction mutuelle, au quel cas la dépendance est positive, soit il y a répulsion, et alors il existe une dépendance positive entre les deux motifs X et $\overline{Y}$ et l'on considère $X \rightarrow \overline{Y}$ d'une part, puis entre $\overline{X}$ et Y et l'on considère $\overline{X} \rightarrow Y$ d'autre part. Dans les deux cas, on aura toujours à considérer une dépendance positive. Décomposons alors la mesure $\mathrm{M_{GK}}$ comme suit :

Définition 15

$$\mathrm{M_{GK}}(X \rightarrow Y) = \begin{cases} {\mathrm{M_{GK}}}^f(X \rightarrow Y), & \textit{si } X \textit{ favorise } Y \\ {\mathrm{M_{GK}}}^d(X \rightarrow Y), & \textit{si } X \textit{ défavorise } Y. \end{cases}$$

Ainsi, c'est surtout la composante favorable ${\mathrm{M_{GK}}}^f$ qui va guider la sémantique de $\mathrm{M_{GK}}$. Entre l'indice pionnier $Confiance$ et $\mathrm{M_{GK}}$, on a la proposition ci-dessous.

Proposition 12 (1) *Aux marginales fixées,* $Confiance$ *et* ${\mathrm{M_{GK}}}^f$ *sont deux fonctions croissantes du nombre d'exemples de règle, cependant* ${\mathrm{M_{GK}}}^f$ *croît plus lentement que* $Confiance$.

(2) $\forall (X \rightarrow Y)$ *telle que* X *favorise* Y, *on a :*
(a) Si leurs extensions sont telles que $X' \subseteq Y'$, *alors* $\mathrm{conf}(X \rightarrow Y) = {\mathrm{M_{GK}}}^f(X \rightarrow Y) = 1 : X \rightarrow Y$ *est dite une règle exacte.*
(b) Si $X' \nsubseteq Y'$, *alors* $0 < {\mathrm{M_{GK}}}^f(X \rightarrow Y) < \mathrm{conf}(X \rightarrow Y) < 1$, *ou encore* $\frac{1-{\mathrm{M_{GK}}}^f}{1-\mathrm{conf}}(X \rightarrow Y) > 1 : X \rightarrow Y$ *est dite une règle approximative.*

En effet : Pour deux motifs X et Y, posons $t = |X' \cap Y'|$. Il vient :
(1)$\mathrm{conf}(X \rightarrow Y) = \frac{t}{n_X} = f_1(t)$ et ${\mathrm{M_{GK}}}^f(X \rightarrow Y) = \frac{\frac{t}{n_X} - P(Y)}{1-P(Y)} = f_2(t)$.
Or ces deux fonctions ont leurs dérivées rangées ainsi :
$0 < f_2'(t) = \frac{1}{n_X(1-P(Y))} < f_1'(t) = \frac{1}{n_X}$.
Il en résulte que $Confiance$ croît plus vite vers la valeur maximale 1 que $\mathrm{M_{GK}}$ dans la zone d'attraction.
Le résultat $2(a)$ est immédiat. Enfin, pour $2(b)$ on vérifie que sous cette hypothèse :
$\frac{1-{\mathrm{M_{GK}}}^f}{1-\mathrm{conf}}(X \rightarrow Y) = \frac{1}{P(\overline{Y'})} > 1$.

Corollaire 3 *La mesure de qualité normalisée* $\mathrm{M_{GK}}$ *a favorablement un pouvoir discriminant plus élevé que* $Confiance$.

De la Proposition 12 et du Corollaire 1, on tire le corollaire suivant.

Corollaire 4 *Pour un même type de règle, la mesure de qualité normalisée* $\mathrm{M_{GK}}$ *produit moins de règles que* $Confiance$.

De plus on a les résultats suivants.

Proposition 13 *(i) Si X favorise Y, nous avons la relation d'équivalence des deux régles contraposées :*

$$\mathrm{M_{GK}}^f(\overline{Y}\rightarrow\overline{X}) = \mathrm{M_{GK}}^f(X\rightarrow Y). \tag{3.7}$$

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

$$\mathrm{M_{GK}}^d(\overline{Y}\rightarrow\overline{X}) = \frac{p(X')p(Y')}{(1-p(X'))(1-p(Y'))}\mathrm{M_{GK}}^d(X\rightarrow Y) \tag{3.8}$$

Ainsi, compte tenu de la remarque faite ci-dessus, on peut considérer que $\mathrm{M_{GK}}$ est favorablement implicative.

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

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

(ii) Si X défavorise Y, nous avons

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

Ce qui démontre les résultats. □

La Proposition 14 ci-dessous montre que la mesure de qualité $\mathrm{M_{GK}}$ est favorablement non symétrique.

Proposition 14 *(i) Si X favorise Y, nous avons la relation :*

$$\mathrm{M_{GK}}^f(Y\rightarrow X) = \frac{1-p(Y')}{1-p(X')}\frac{p(X')}{p(Y')}\mathrm{M_{GK}}^f(X\rightarrow Y). \tag{3.9}$$

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

$$\mathrm{M_{GK}}^d(Y \to X) = \mathrm{M_{GK}}^d(X \to Y) \tag{3.10}$$

Pour ce qui concerne les règles négatives à droite, nous avons :

Proposition 15 *Soient X et Y deux motifs positifs.*
On a l'égalité et l'équivalence suivantes :

$$\mathrm{M_{GK}}^f(X \to \overline{Y}) = -\mathrm{M_{GK}}^d(X \to Y). \tag{3.11}$$

Et $\forall \alpha \in]0,1[$, ona :
$$(-1 < \mathrm{M_{GK}}^d(X \to Y) < -\alpha \Longleftrightarrow \alpha < \mathrm{M_{GK}}^f(X \to \overline{Y}) < 1)$$

En effet :

$$\begin{array}{rcl} \mathrm{M_{GK}}^f(X \to \overline{Y}) & = & \frac{P(\overline{Y}'|X') - P(\overline{Y}')}{1 - P(\overline{Y}')} \\ & = & \frac{1 - P(Y'|X') - 1 + P(Y')}{1 - 1 + P(Y')} \\ & = & \frac{-P(Y'|X') + P(Y')}{P(Y')} \\ & = & -\mathrm{M_{GK}}^d(X \to Y) \end{array}$$

Ce qui démontre l'égalité de la proposition 15. □

De cette proposition 15 résulte que plus le degré de quasi-incompatibilité entre deux motifs est élevé, plus la qualité de la règle négative à droite correspondante est favorablement meilleure. Alors que l'équivalence permet d'élaguer directement toute candidate règle négative à droite dont la valeur $\mathrm{M_{GK}}$ de la règle positive associée est négative et située dans l'intervalle $[-\alpha, 0[$, pour un seuil α fixé dans $]0,1[$. Au sujet des règles négatives à gauche, on a la proposition 16 qui suit.

Proposition 16 *Pour deux motifs positifs X et Y, on a les égalités suivantes.*

(1) Si X défavorise Y (donc, X favorise $\overline{Y}$ et aussi $\overline{X}$ favorise Y), alors :

$$\mathrm{M_{GK}}^f(\overline{X} \to Y) = \frac{P(X')}{1 - P(X')} \frac{P(Y')}{1 - P(Y')} \mathrm{M_{GK}}^f(X \to \overline{Y}) \tag{3.12}$$

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

$$\mathrm{M_{GK}}^d(\overline{X} \to Y) = \frac{P(X')}{(1 - P(X'))} \frac{1 - P(Y')}{P(Y')} \mathrm{M_{GK}}^d(X \to \overline{Y}) \tag{3.13}$$

En effet :

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

Pour le point (2) de cette proposition, on a :

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

Ce qui démontre les résultats. □

À partir des deux propositions 15 et 16 précédentes, on tire les relations entre une règle négative à gauche et la règle positive correspondante.

Corollaire 5 *Pour deux motifs positifs X et Y, on a les égalités suivantes.*
(1) Si X défavorise Y (donc, X favorise $\overline{Y}$ et aussi $\overline{X}$ favorise Y), alors :

$$\mathrm{M_{GK}}^f(\overline{X}\rightarrow Y) = -\frac{p(X')}{1-P(X')}\frac{P(Y')}{1-P(Y')}\mathrm{M_{GK}}^d(X\rightarrow Y) \qquad (3.14)$$

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

$$\mathrm{M_{GK}}^d(\overline{X}\rightarrow Y) = -\frac{P(X')}{(1-P(X'))}\frac{1-P(Y')}{P(Y')}\mathrm{M_{GK}}^f(X\rightarrow Y) \qquad (3.15)$$

Partant maintenant de ce corollaire, on obtient aisément une relation (cf.corollaire 6) entre seuils permettant de sélectionner des règles négatives à gauche intéressantes ou d'élaguer celles non intéressantes.

Corollaire 6 *Pour deux motifs positifs X et Y, on a les égalités suivantes. Si X défavorise Y (donc, X favorise $\overline{Y}$ et aussi $\overline{X}$ favorise Y), alors pour un seuil $\alpha \in]0,1[$ fixé :*

$$-1 < \mathrm{M_{GK}}^d(X\rightarrow Y) < -\alpha \Longleftrightarrow \frac{p(X')}{1-P(X')}\frac{P(Y')}{1-P(Y')}\alpha < \mathrm{M_{GK}}^f(\overline{X}\rightarrow Y) < \frac{p(X')}{1-P(X')}\frac{P(Y')}{1-P(Y')}$$

Proposition 17 *Aux motifs emboîtés sont associées des règles d'association exactes ou implications totales.*
De plus :

(i) Si $X \subseteq Y \subseteq Z \subseteq \mathcal{A}$ *alors*

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

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

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

Par conséquent, $\mathrm{M_{GK}}$ est multiplicative sur une chaîne du treillis de motifs. Ce qui apparaît intéressant dans la pratique, sur le plan d'élagage grâce à la transitivité axiale. Par ailleurs, il est à remarquer aussi que cette propriété demeure pour toute mesure de qualité normalisée continue. Le corollaire suivant découle de la Proposition 17 ci-dessus.
Démonstration : Il suffit de démontrer le point (i). Le point (ii) s'obtient facilement à l'aide d'un raisonnement par récurrence. En passant aux extensions, les inclusions $X \subseteq Y \subseteq Z$ impliquent $Z' \subseteq Y' \subseteq X'$.
Donc $P(Y'|X') = \frac{P(X' \cap Y')}{P(X')} = \frac{P(Y')}{P(X')} \geq P(Y')$. Donc X favorise$Y$.
De façon analogue, X favorise$Z$ et Y favorise$Z$.
Ainsi, les trois motifs X, Y et Z se favorisent deux à deux. Par conséquent, on a :

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

Nous avons donc,

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

Ce qui démontre le résultat. □

Corollaire 7 *Soient* $X_1, X_2, ..., X_i, X_{i+1}, ..., X_p$ *des motifs tels que* $X_1 \subseteq X_2 \subseteq ... \subseteq X_i \subseteq X_{i+1} \subseteq ... \subseteq X_p$.

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

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

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

Une mesure de qualité est dite "mesure de *déviation d'équilibre*" si elle prend une valeur constante quand le nombre d'exemples et de contre-exemples de la règle sont égaux [BGBG05]. La Proposition 18 ci-dessous montre que la mesure de qualité $\mathrm{M_{GK}}$ est une mesure de déviation d'équilibre pour n suffisamment grand, *i.e.*, pour un grand volume de données.

Proposition 18 ***(Situation de référence à l'équilibre)*** *[TRD05, DRT07]*

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

En effet :

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

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

3.3.2 Seuils de signification de $\mathrm{M_{GK}}$

Par ailleurs, par la considération du tableau de contingence obtenu par le croisement des deux motifs, à l'instar de la relation fonctionnelle existant entre l'indice d'implication de Gras et la statistique de χ^2 de Pearson à un degé de liberté établie dans (voir [Tot94] ou [GSB+96, pp 43-47], sachant que pour un contexte binaire, on a l'avantage de ne pas être soumis à une condition de normalité, la proposition 19 ci-dessous montre qu'il est facile d'élaborer les abaques des valeurs critiques pour la signification d'une règle d'association valide selon $\mathrm{M_{GK}}$.
Pour deux motifs U and V du contexte $\mathbb{K}$ on a :

Proposition 19

$$M_{GK}(U \rightarrow V) = \begin{cases} \sqrt{\frac{1}{n} \frac{n - n_U}{n_U} \frac{n_V}{n - n_V} \chi^2}, & \text{si } U \text{ fav. } V \text{ ou } U \text{ et } V \text{ sont indépdts} \\ -\sqrt{\frac{1}{n} \frac{n - n_U}{n_U} \frac{n - n_V}{n_V} \chi^2}, & \text{si } U \text{ défav. } V \text{ ou } U \text{ et } V \text{ sont indépdts} \end{cases}$$

$U \diagdown V$	V	$\overline{V}$	
U	3000	2000	5000
$\overline{U}$	2500	2500	5000
	5500	4500	10000

(1) Dependance positive
$\chi^2 = 101$, $M_{GK} = +0.11$

$U \diagdown V$	V	$\overline{V}$	
U	1000	3000	4000
$\overline{U}$	4500	1500	6000
	5500	4500	10000

(2) Dépendance négative
$\chi^2 = 2424$, $M_{GK} = -0.54$

$U \diagdown V$	V	$\overline{V}$	
U	2200	1800	4000
$\overline{U}$	3300	2700	6000
	5500	4500	10000

(3) Indépendance
$\chi^2 = 0 = M_{GK}$

$U \diagdown V$	V	$\overline{V}$	
U	0	2000	2000
$\overline{U}$	6000	2000	8000
	6000	4000	10000

(4) Incompatibility
$\chi^2 = 3750$, $M_{GK} = -1$

$U \diagdown V$	V	$\overline{V}$	
U	3000	0	3000
$\overline{U}$	3000	4000	7000
	6000	4000	10000

(5) Logical implication
$\chi^2 = 2857$, $M_{GK} = 1$

FIG. 3.2 – Comparaison de M_{GK} et χ^2 sur cinq situations de référence

La figure 3.2 fournit plus d'éclairage sur les modes d'évaluation du degré de dépendance entre deux motifs, dans cinq situations de référence : dépendance positive (Fig. 3.2 (1)), dépendance négative (Fig. 3.2 (2)), indépendance (Fig. 3.2 (3)), incompatibilité (Fig. 3.2 (4)), et l'implication logique (Fig. 3.2 (5)). Il apparaît que, contrairement à χ^2, la mesure $\mathrm{M_{GK}}$ calcule le degré de dépendance orientée sur une échelle limitée dans l'intervalle $[-1, +1]$, en plus de son orientation : Par exemple, dans le tableau (Fig. 3.2 (2)) la très significative dépendance entre les deux motifs revélée par χ^2 est en fait une dépendance négative et comme $\mathrm{M_{GK}}^f(U \rightarrow \overline{V}) = 0,1$, donc très faible devant $\mathrm{M_{GK}}^f(\overline{U} \rightarrow V) = 4/9 = 0,444,1$, seule la règle négative à gauche $\overline{U} \rightarrow V$ est significativement valide.

Notons également qu'ici la $\mathrm{conf}(\overline{U} \rightarrow V) = 0,75$ est suffisamment élevée aussi, alors que son évaluation selon $\mathrm{M_{GK}}$ reste relativement faible (donc à taux de saturation$(\mathrm{M_{GK}}) = \frac{1-\mathrm{M_{GK}}}{1-0} = \frac{5}{9} \simeq 0.556 > (\frac{1-\mathrm{conf}}{1-0} = \frac{1}{4} = 0.25)$).

En fait, il est immédiat qu'en cas de lien positif non totalement implicatif, $0 < \frac{\mathrm{M_{GK}}^f}{\mathrm{conf}} = Zang^f < 1$. Ce qui conduit bien à affirmer que $\mathrm{M_{GK}}$ est plus discriminante que la confiance.

3.3.3 M_{GK} et coéfficient de corrélation linéaire

, Considérons deux motifs X et Y du contexte. Leurs extensions respectives X' et Y' sont deux événements de $\mathcal{P}(\mathcal{O})$. Soient $\mathbf{1}_{\mathbf{X}'}$ et $\mathbf{1}_{\mathbf{Y}'}$ les indicatrices respectives de ces derniers. L'espérance mathématique de $\mathbf{1}_{\mathbf{X}'}$ est $E(\mathbf{1}_{\mathbf{X}'}) = P(X')$. De même, celle de $\mathbf{1}_{\mathbf{Y}'}$ est $E(\mathbf{1}_{\mathbf{Y}'}) = P(Y')$.
Les variances respectives de ces variables indicatrices sont : $\sigma^2_{\mathbf{1}_{\mathbf{X}'}} = P(X')P(\overline{X'})$

et $\sigma^2_{1_b f 1_{Y'}} = P(Y)P(\overline{Y})$. La covariance de ces variables aléatoires indicatrices est : $cov(\mathbf{1_{X'}}, \mathbf{1_{Y'}}) = E(\mathbf{1_{X'}}\mathbf{1_{Y'}}) - E(\mathbf{1_{X'}}E(\mathbf{1_{X'}} = P(X' \cap Y') - P(X')P(Y')$.
Comme

$$\mathrm{M_{GK}}(X \to Y) = \begin{cases} \frac{P(Y' \cap X') - P(X')P(Y')}{P(X')(1-P(Y'))}, & si \text{ X } favorise \text{ Y } ou \text{ X } et \text{ Y } sont \text{ indépendants} \\ \\ \frac{P(Y' \cap X') - P(X')P(V')}{P(X')P(Y')}, & si \text{ X défavorise Y } ou \text{ X } et \text{ Y indépendants} \end{cases}$$

D'où les relations cléfs entre la mesure $\mathrm{M_{GK}}$ et la covariance, puis le coéfficient de corélation linéaire ρ :

$$\mathrm{M_{GK}}(X \to Y) = \begin{cases} \frac{Cov(I_{X'}, I_{Y'})}{P(X')(1-P(Y'))}, & \text{si } X \text{ favorise } Y \text{ ou } X \text{ et } Y \text{ sont indépendants} \\ \\ \frac{Cov(I_{X'}, I_{Y'})}{P(X')P(Y')}, & \text{si } X \text{ défavorise } Y \text{ ou } X \text{ et } Y \text{ indépendants} \end{cases}$$

$$\mathrm{M_{GK}}(X \to Y) = \begin{cases} \rho(\mathbf{1_{X'}}, \mathbf{1_{Y'}}) \sqrt{\frac{P(\overline{X}')}{P(X')} \frac{P(Y')}{P(\overline{Y}')}}, & \text{si } X \text{ favorise } Y \text{ ou } X \text{ et } Y \text{ sont indépendants} \\ \\ \rho(1_{X'}, 1_{Y'}) \sqrt{\frac{P(X')}{P(\overline{X}')} \frac{P(Y')}{P(\overline{Y}')}}, & \text{si } X \text{ défavorise } Y \text{ ou } X \text{ et } Y \text{ indépendants} \end{cases}$$

$$\mathrm{M_{GK}}(X \to Y) = \begin{cases} \rho(X,Y) \sqrt{\frac{n_{\overline{X}}}{n_X} \frac{n_Y}{n_{\overline{Y}}}}, & \text{si } X \text{ favorise } Y \text{ ou } X \text{ et } Y \text{ sont indépendants} \\ \\ \rho(X,Y) \sqrt{\frac{n_X}{n_{\overline{X}}} \frac{n_Y}{n_{\overline{Y}}}}, & \text{si } X \text{ défavorise } Y \text{ ou } X \text{ et } Y \text{ indépendants} \end{cases}$$

Et réciproquement :
$\rho(X,Y) = \sqrt{\frac{P(X')P(\overline{Y'})}{P(\overline{X'})P(Y')}} \mathrm{M_{GK}}^f(X \to Y)\mathbf{1_f}(X \to Y) + \sqrt{\frac{P(X')P(Y')}{P(\overline{X'})P(\overline{Y'})}} \mathrm{M_{GK}}^d(X \to Y)\mathbf{1_d}(X \to Y)$.
Ainsi, la mesure $\mathrm{M_{GK}}$ et le coéfficient de corrélation linéaire ρ, c'est-à-dire le $\phi - coef ficient$, ont le même signe. Plus précisement, la mesure probabiliste implicative orientée $\mathrm{M_{GK}}$ intègre la corrélation et oriente cette dépendance eu égard à sa propriété non symétrique. Notons que cette propriété, comme la sémantique d'implication statistique, est partagée par toute les mesures probabilistes de la qualité normalisées continues.
D'où les diverses manières d'interpréter une règle $\mathrm{M_{GK}}$-valide, ou valides selon une mesure de qualité normaliée continue quelconque, de la même manière que pour le coéfficient de corrélation linéaire $\rho(X,Y)$ à la différence de la non symétrie cette fois :
- En terme de similarité dynamique : une dynamique vers l'identité de la variable prémisse sur la variable conséquant (closeness to identity) ;
- En terme de moyenne géométrique (averaging the slopes)
- En génétique et en démographie : vers une probabilité d'une descendance commune (Probability of common descent) ;
- En recherche didactique de disciplines : vers la conjecture des conditions nécessaires à la réussite d'un concept précis, soit vers une mise en évidence d'une taxonomie des objectifs cognitifs.

3.4 Génération de Base composite des règles d'association $\mathrm{M_{GK}}$-valides

Nous savons que les règles positives au sens de notre approche ne suffisent pas pour couvrir tous les besoins dans un contexte de fouille de données. Il faut aussi des règles n'egatives à gauche et des règles négatives à droite. Or les bases des règles existantes ne concerne que les règles positives (cf. par exemples algorithme 1 et [Pas00]), et ce, selon les mesures support et confiance exclusivement. L'algorithme de génération de la base de Guigues-Duquenne présenté dans l'algorithme 1 ci-dessous est celui proposé dans [Pas00]. Cet algorithme suppose que les ensembles des motifs fréquents et fermés fréquents sont déjà calculés. Il existe plusieurs algorithmes de génération des motifs fermés dans la littérature ; citons entre autres : CLOSE [PBTL99], CLOSET [PHM00], CHARM [ZH99], TITANIC [STB+02], PRINCE [HYS05]. Les notations utilisées dans l'algorithm 1 sont présentées dans le Tableau 3.4.

$\mathcal{F}_i$	Les i-motifs fréquents. Chaque élément de $\mathcal{F}_i$ possède deux champs : motifs et Support.
$\mathcal{FC}_i$	Les i-motifs fermés fréquents. Chaque élément de $\mathcal{FC}_i$ possède deux champs : motif et Support.
$\mathcal{CF}_k$	Les i-motifs critiques potentiels.
BGD	Base de Gugues-Duquenne.
Cumul	Union des fermetures des sous-ensembles critiques. du i-motif critique fréquent candidat X considéré
k	Taille maximale de motifs fermés fréquents.

Tab. 3.4 – Notations utilisées dans l'algorithme 1

Algorithm 1
Entrée : $\mathcal{F}_i$ ensemble des i-motifs fréquents ; $\mathcal{FC}_i$ ensemble de i-motifs fermés fréquents. **Sortie :** $\mathcal{F} : BGD$ la base de Guigues-Duquenne ;

1: $BGD \leftarrow \{\}$;
2: **if** $(\varphi(\emptyset) \neq \emptyset)$ **then**
3: $\quad BGD \leftarrow BGD \cup \{\emptyset \rightarrow \varphi(\emptyset)\}$
4: **end if**
5: **for all** $\mathcal{F}_i$, pour $i < k$ **do**
6: $\quad CPF_i \leftarrow \mathcal{F}_i \setminus \mathcal{FC}_i)$;
7: $\quad$ **for all** $X \in CPF_i$ **do**
8: $\quad\quad Cumul \leftarrow \emptyset$;
9: $\quad\quad$ **for all** $(C \rightarrow \varphi(C) \setminus C) \in BGD$ **do**
10: $\quad\quad\quad$ **if** $C \subseteq X$ **then**
11: $\quad\quad\quad\quad cumul \leftarrow cumul \cup \varphi(C)$;
12: $\quad\quad\quad$ **end if**

13: **end for**
14: **if** *cumul* $\subset X$ **then**
15: $BGD \leftarrow BGD \cup \{X \rightarrow \varphi(X) \setminus X\}$;
16: **end if**
17: **end for**
18: **end for**
19: Retourner BGD ;

L'algorithme commence par initialiser l'ensemble BGD avec l'ensemble vide (ligne 1). Il détermine ensuite si l'ensemble $\emptyset$ est fermé ou non (s'il n'est pas fermé, il necessairement critique). Si $\emptyset$ est critique, la règle $\emptyset \rightarrow \varphi(\emptyset)$ est inserée dans BGD (ligne 3). Ensuite, la boucle dans les étapes (lignes 5-18) constuit la base de Guigues-Duquenne de façon itérative. Durant une itération i, l'ensemble CFP_i de i-motifs critiques fréquents candidats est initilalisé avec les i-motifs fréquents $X \in \mathcal{F}_i$, qui ne sont pas des i-motifs fermés fréquents (ligne 6). Ensuite, chacun des i-motifs critiques fréquents candidats C est examiné afin de déterminer s'il est critique (ligne 7-16). Pour cela, l'union des fermetures des sous-ensembles critiques du motif X est calculée dans le motif *cumul* (lignes 8-13). Ces sous-ensembles critiques sont les antécédents C des règles dans la base de Guigues-Duquenne. Si le motif *cumul* est inclus dans le motif X alors X est un motif critique fréquent et la règle $X \rightarrow \varphi(X) \setminus X$ est insérée dans la base BGD (lignes 14-16). La boucle s'arrête lorsque l'ensemble les motifs fréquents a été considéré et l'ensemble BGD retourné par l'algorithme contient toutes les règles de la base de Guigues-Duquenne.

D'où notre intérêt pour trouver un moyen de génération de bases pour les deux types de règles, et ce selon une autre mesure de qualité normalisée qui élague systématiquement entre autres les règles 'a conséquent prémisse indépendfants : avec toutes les qualités que jouit la mesure normalisée implicative, notre choix s'impose sur $\mathrm{M_{GK}}$. Dans cette partie nous proposons des axiomes d'inférence des bases des règles d'association $\mathrm{M_{GK}}$-intéressantes (ou valides) [FDT06b, DFT06]. Tout d'abord, comme pour tous motifs X, Y ona l'équivalence $\mathrm{M_{GK}}(X \rightarrow Y) = 1 \Longleftrightarrow \mathrm{conf}(X \rightarrow Y) = 1$, la base pour les règles $\mathrm{M_{GK}}$-exactes coïncide avec la base bien connue de Guigue-Duquenne-Luxenburger (soit BRPE). Il ne reste plus qu'à compléter celle-ci par une base des règles négatives exactes (BRNE) et les trois autres types de base des règles approximatives (i.e. non exactes), 'a savoir : base des règles positives approximatives (BRPA), base des règles approximatives négatives à gauche (BRNAG), et base des règles approximatives négatives à droite (BRNAD), au sens de la mesure $\mathrm{M_{GK}}$. Or il est immédiat que :
pour tous motifs positifs X et Y, $\mathrm{M_{GK}}(\overline{X} \rightarrow Y) = \frac{p(X')}{1-p(X')} \frac{p(Y')}{1-p(Y')} \mathrm{M_{GK}}(X \rightarrow \overline{Y})$.
Il en résulte que les règles négatives à gauche peuvent être dérivées par celles négatives à droitre et vice-versa. Grâce à ces deux propriétés, nous ne considérons dans la suite que deux types des règles d'association, à savoir les règles positives et celles négatives à droites que nous appelerons tout sim-

plement règles négatives. Nous donnons ci-dessous le résulat obtenu pour chacun des cas ([Fen07, FDT06b]).

3.4.0.1 Base pour les règles négatives exactes (BRNE)

De l'équivalence $\mathrm{M_{GK}}(X \rightarrow \overline{Y}) = 1 \Longleftrightarrow \mathrm{supp}(X \rightarrow Y) = 0$, on aboutit àux deux axiomes d'inférence sur les règles négatives exactes suivants :

Lemme 4 *: Axiome d'inférence pour règle négatives exactes Pour tous* X*,* Y *et* Z *:*
(NE1) $X \rightarrow \overline{Y}$ *et* $\mathrm{supp}(Y \cup Z) > 0$ *impliquent* $X \rightarrow \overline{Y \cup Z}$ *;*
(NE2) $X \rightarrow \overline{Y}, Z \subset X$ *et* $\mathrm{supp}(Z \cup Y) = 0$ *impliquent* $Z \rightarrow \overline{Y}$*.*

Proposition 20 *Les axiomes d'inférence* $NE1$ *et* $NE2$ *sont corrects pour les règles négatives exactes, i.e., toute règle d'association déduite, par application de (NE1) et (NE2), à partir d'une règle d'association négative exacte est négative exacte.*

Démonstration Nous montrons d'abord que $(NE1)$ est correct. Soit $X \rightarrow \overline{Y}$ une règle négative exacte, *i.e.*, $\mathrm{M_{GK}}(X \rightarrow \overline{Y}) = 1$. Or dans ce cas $\mathrm{supp}(X \cup Y) = 0$. Donc, pour tout motif Z, on a $\mathrm{supp}(X \cup (Y \cup Z)) = \mathrm{supp}(X \cup Y \cup Z) = 0$. Par ailleurs, si Z tel que $\mathrm{supp}(Z \cup Y) > 0$. D'où $\mathrm{M_{GK}}(X \rightarrow \overline{Y \cup Z}) = 1$. Ce qui démontre la correction de $(NE1)$.
Maintenant, montrons que $(NE2)$ est correct. Soit $X \rightarrow \overline{Y}$ une règles négative exacte, *i.e.*, $\mathrm{M_{GK}}(X \rightarrow \overline{Y}) = 1$. Donc pour tout motif Ztel que $Z \subset X$ on a $\mathrm{supp}(X) > 0$. Ainsi, si $\mathrm{supp}(Z \cup Y) = 0$, alors, par la Proposition **??**, on a $\mathrm{M_{GK}}(Z \rightarrow \overline{Y}) = 1$. Ce qui démontre que $Z \rightarrow \overline{Y}$ est une règle négative exacte. □ Considérons la bordure positive de l'ensemble des motifs de Support non nul $Bd^{+}(0)$ [MT97] définie par :

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

Remarque 9 *Notons que la bordure positive* $Bd^{+}(0)$ *est l'ensemble des motifs maximaux de Support non nul. Elle est identique à l'ensemble des motifs fermés maximaux de Support non nul [PBTL99].*

Nous caractérisons maintenant la base que nous proposons pour l'ensemble des règles négatives exactes $\mathrm{M_{GK}}$-valides.

Thèorème 4 *[FDT06b, DFT06] L'ensemble* $BRNE$ *défini par :*

$$BRNE = \{X \rightarrow \{\overline{x}\} : X \in Bd^{+}(0) \text{ et } x \notin X\}.$$

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

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

Montrons maintenant que l'ensemble BNE est minimal. Soit $X \rightarrow \overline{x}$ un élément de BRNE et soit BRNE$' =$ BRNE $- \{X \rightarrow \overline{x}\}$. Montrons que la règle $X \rightarrow \overline{x}$ ne peut pas être dérivée de BRNE$'$ par application de (NE1) et (NE2). En effet, la règle $X \rightarrow \overline{x}$ ne peut pas être dérivée d'une règle $X \rightarrow \overline{Y}$ par application de (NE1) car cela impliquerait nécessairement $Y \subset \{x\}$. D'autre part, la règle $X \rightarrow \overline{x}$ ne peut pas être dérivée d'une autre règle $Z \rightarrow \overline{x}$ par application (NE2). En effet, cela impliquerait que $X \subset Z$ donc $supp(Z) = 0$ puisque $X \in Bd^+(0)$. D'où, la règle $X \rightarrow \overline{x}$ ne peut pas être dérivée d'une règle de BNE$'$, ce qui démontre la minimalité de BNE. □

Exemple 2 *La base $BRNE$, pour les règles négatives exactes, extraite du contexte du Tableau 3 est $BRNE = \{ACD \rightarrow \overline{B}, ACD \rightarrow \overline{E}, ABCE \rightarrow \overline{D}\}$.*

La règle $ABCE \rightarrow \overline{D}$ est une règle de la base $BRNE$. Par application des axiomes (NE1) et (NE2), nous pouvons dériver à partir de cette règles les onze règles : $ABCE \rightarrow \overline{AD}$, $ABCE \rightarrow \overline{CD}$, $ABE \rightarrow \overline{ACD}$, $BE \rightarrow \overline{AD}$, $E \rightarrow \overline{AD}$, $B \rightarrow \overline{AD}$, $E \rightarrow \overline{CD}$, $B \rightarrow \overline{AD}$, $E \rightarrow \overline{ACD}$, $B \rightarrow \overline{ACD}$.

Remarque 10
- *Comme la bordure positive $Bd^+(0)$ est identique à l'ensemble des motifs φ-fermés maximaux de Support strictement positif, donc la base $BRNE$, pour les règles négatives exactes, est exprimée en terme de de l'opérateur de fermeture φ.*
- *Si la règle $X \rightarrow \overline{Y}$ est exacte alors la règle $Y \rightarrow \overline{X}$ l'est aussi, et réciproquement. Toutefois, ces deux règles n'ont pas toujours le même degré d'informativité. En effet, si $|X_1| > |X_2| > |Y_1| > |Y_2|$, alors la règle $X_2 \rightarrow \overline{Y}_2$ est la plus informative que toutes autres règles négatives exactes M_{GK}-valides combinant les motifs X_1, X_2, Y_1, Y_2.*

Dans [FDT07, DdRFT07], nous proposons un algorithme de génération de la base $BRNE$ pour les règles négatives exactes M_{GK}-valides extraite d'un contexte de la fouille de données $\mathbb{K}$. Le pseudo-code de l'algorithme générant la base $BRNE$ est présenté dans l'algorithme 2. Le présent algorithme suppose que la bordure positive $Bd^+(0)$ est déjà trouvée. Il existe dans la littérature différents algorithmes permettant de générer la bordure positive ou les fermés maximaux [LK98, Bay98, ZPOL97]. L'algorithme 2 commence par initialiser l'ensemble $BRNE$ à l'ensemble vide (ligne 1). Chaque élément X de l'ensemble $Bd^+(0)$ est examiné successivement (lignes 2 à 6). Pour

chaque attribut $x \notin X$, la règle $X \rightarrow \overline{x}$ est insérée dans $BRNE$ (lignes 3 à 5).

Algorithm 2 (Base Négative Exacte)
Entrée : $Bd^{+}(0)$.
Sortie : BRNE.

1: $BRNE \leftarrow \{\}$
2: **for all** $(X \in Bd^{+}(0))$ **do**
3: **for all** $x \notin X$ **do**
4: $BRNE \leftarrow BRNE \cup \{X \rightarrow \overline{x}\}$
5: **end for**
6: **end for**
7: Retourner $BRNE$

3.4.1 Base pour les règles positives approximatives

Fixons un seuil $\alpha \in]0,1[$. Le résultat de la Proposition 21 ci-dessous caractérise les règles positives approximatives $(\mathrm{M_{GK}}, \alpha)$-valides en fonction de leurs Confiances respectives.

Proposition 21 *Si X et Y sont deux motifs tels que X favorise Y alors, on a l'équivalence : $\alpha \leq M_{GK}(X \rightarrow Y) < 1 \Longleftrightarrow \mathrm{supp}(Y)(1-\alpha)+\alpha \leq \mathrm{conf}(X \rightarrow Y) < 1$.*

La preuve est immédiate.

Considérons l'opérateur de fermeture φ de la famille $\mathcal{P}(\mathcal{A})$ évoqué dans le chapitre qui précède. De cette proposition 21 nous sommes conduits à considérer l'axiome d'inférence (PA) ci-dessous :

Lemme 5 *: Axiome d'inférence sur les règles positives approximatives.*

$(AIRPA)$ si $X \rightarrow Y$ et Z, T sont tels que $\varphi(X) = \varphi(Z)$ et $\varphi(Y) = \varphi(T)$, alors $Z \rightarrow T$.

Les deux lemmes suivants seront utiles pour la démonstration de la Proposition 22 et le Théorème 5.

Le Lemme 6 montre que le support d'un motif est égal au Support de sa fermeture [PBTL99].

Lemme 6 *Pour tout motif X, on a :* $\mathrm{supp}(\varphi(X)) = \mathrm{supp}(X)$.

Le Lemme 7 est une caractérisation des opérateurs de fermeture utilisant une propriété dite d'indépendance de chemins [Plo73].

Lemme 7 *Une application extensive ϕ sur $\mathcal{P}(\mathcal{A})$, i.e. $X \subseteq \phi(X)$, est un opérateur de fermeture sur $\mathcal{P}(\mathcal{A})$ si et seulement si elle vérifie la propriété $\phi(X \cup Y) = \phi(\phi(X) \cup \phi(Y))$, pour tous $X, Y \in \mathcal{P}(\mathcal{A})$.*

Proposition 22 *L'axiome d'inférence (AIRPA) est correct pour les règles négatives approximatives* $(\mathrm{M_{GK}}, \alpha)$*-valides, i.e., toute règle d'association déduite par application de (AIRPA) à partir d'une règle positive approximative* $(\mathrm{M_{GK}}, \alpha)$*-valide est positive approximative* $(\mathrm{M_{GK}}, \alpha)$*-valide.*

Démonstration : Soit $X \rightarrow Y$ une règle positive approximative $(\mathrm{M_{GK}}, \alpha)$-valide règle d'association, *i.e.*, $\alpha \leq \mathrm{M_{GK}}(X \rightarrow Y) < 1$. Alors, par la Proposition 21, $\mathrm{supp}(Y)(1-\alpha)+\alpha \leq conf(X \rightarrow Y) < 1$. Soient Z et T deux motifs tels que $\varphi(X) = \varphi(Z)$ et $\varphi(Y) = \varphi(T)$. Alors, par Lemmes 6 et 7, $\mathrm{supp}(X \cup Y) = \mathrm{supp}(\varphi(X \cup Y)) = \mathrm{supp}(\varphi(\varphi(X) \cup \varphi(Y))) = \mathrm{supp}(\varphi(\varphi(Z) \cup \varphi(T))) = \mathrm{supp}(\varphi(Z \cup T)) = \mathrm{supp}(Z \cup T)$. Par ailleurs, $\mathrm{conf}(Z \rightarrow T) = \mathrm{conf}(X \rightarrow Y)$ donc $\mathrm{supp}(T)(1-\alpha)+\alpha \leq \mathrm{conf}(Z \rightarrow T) < 1$. Alors, pour la même raison que ci-dessus, $\alpha \leq \mathrm{M_{GK}}(Z \rightarrow T) < 1$, ce qui démontre que $Z \rightarrow T$ est approximative $(\mathrm{M_{GK}}, \alpha)$-valide. □

Par ailleurs, nous avons le résultat suivant :

Thèorème 5 *[FDT06b, DFT06] L'ensemble* $BPA(\alpha)$ *défini par*

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

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

Démonstration : Nous commençons par montrer que toute règle positive approximative $(\mathrm{M_{GK}}, \alpha)$-valide peut être dérivée de $\mathrm{BPA}(\alpha)$ par application de l'axiome (PA). Soit $X \rightarrow Y$ une règle positive approximative $(\mathrm{M_{GK}}, \alpha)$-valide. Alors, par la Proposition 21, $\mathrm{supp}(Y)(1-\alpha)+\alpha \leq \mathrm{conf}(X \rightarrow Y) < 1$. Considérons les deux motifs φ-fermés $Z = \varphi(X)$ et $T = \varphi(Y)$. D'une part, par le Lemme 6, $\mathrm{conf}(\varphi(X) \rightarrow \varphi(Y)) = \mathrm{supp}(\varphi(X) \cup \varphi(Y))/\mathrm{supp}(\varphi(X)) = \mathrm{supp}(\varphi(\varphi(X) \cup \varphi(Y)))/\mathrm{supp}(\varphi(X))$ qui, par le Lemme 7, est égale à $\mathrm{supp}(\varphi(X \cup Y))/\varphi(X)$ et qui, encore par le Lemme 6, est égale à $\mathrm{supp}(X \cup Y)/sup(X) = \mathrm{conf}(X \rightarrow Y)$. D'autre part, par le Lemme 6, $\mathrm{supp}(\varphi(Y) = \mathrm{supp}(Y))$, donc $\mathrm{supp}(\varphi(Y))(1-\alpha)+\alpha) \leq \mathrm{conf}(\varphi(X) \rightarrow \varphi(Y)) < 1$. Donc, par la Proposition 21, $0 < \mathrm{M_{GK}}(\varphi(X) \rightarrow \varphi(Y)) < 1$ alors $\varphi(X) \rightarrow \varphi(Y)$ est un élément de $\mathrm{BPA}(\alpha)$. Par ailleurs, l'application de (PA) à $Z \rightarrow T$ donne la règle $X \rightarrow Y$.

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

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

Alors $X_1 = \varphi(X_1) = \cdots = \varphi(X_n) = \varphi(X) = X$ et $Y_1 = \varphi(Y_1) = \cdots = \varphi(Y_n) = \varphi(Y) = Y$ avec $X_1 \rightarrow Y_1 \in \mathrm{BPA}'$, ce qui contredit le fait que $X \rightarrow Y \notin$

BPA′(α). Donc, $X \rightarrow Y$ ne peut pas être dérivée de $BPA'(\alpha)$, démontrant la minimalité de BPA(α). □

Remarque 11 *Dans la pratique, la base considérée est la restriction de la base ainsi définie sur l'ensemble de motifs fréquents.* $BPA(\alpha) = \{X \rightarrow Y : \varphi(X) = X, \varphi(Y) = Y, X, Y$ *fréquents* $\text{supp}(Y)(1-\alpha) + \alpha \leq \text{conf}(X \rightarrow Y) < 1\}$.

Exemple 3

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

TAB. 3.5 – Contexte de la fouille de données

La base positive approximative BPA, pour les règles d'association positives approximatives M_{GK}*-valides, extraite du contexte du Tableau 3 (avec* $\text{minsupp} = \frac{2}{6}$ *min*$M_{GK} = \frac{2}{6}$*) est* $BPA(\frac{2}{6}) = \{AC \rightarrow ABCE, BE \rightarrow BCE\}$. $AC \rightarrow ABCE$ *est une règle de la base* $BPA(\frac{2}{6})$*. Par application de l'axiome d'inférence (PA), nous pouvons dériver les neuf règles d'association* $A \rightarrow AB$, $A \rightarrow AE$, $A \rightarrow ABC$,$A \rightarrow ACE$, $A \rightarrow ABCE$ $AC \rightarrow AB$, $AC \rightarrow AE$, $AC \rightarrow ACE$, $AC \rightarrow ABCE$.

3.4.2 Base pour les règles négatives approximatives

Rappelons qu'une règle valide au sens de M_{GK} est nécessairement une règle dont la prémisse favorise le conséquent. Aussi, malgré la mutualité qualitative d'une dépendance négative, par souci de cohérence, nous allons considérer seulement les règles négatives à droite du type $X \rightarrow \overline{Y}$, car X favorise $\overline{Y}$ lorsque X défavorise Y.

Caractérisons d'abord les règles négatives approximatives (M_{GK}, α)-valides, i.e. les règles $X \rightarrow \overline{Y}$ telles que $\alpha \leq M_{GK}(X \rightarrow \overline{Y}) < 1$. Nous avons le ésultat immédiat suivant :

Lemme 8 *Soient X et Y deux motifs tels que X défavorise Y, i.e., X favorise* $\overline{Y}$*. Alors* $\alpha \leq M_{GK}(X \rightarrow \overline{Y}) < 1$ *si et seulement si* $0 < \text{conf}(X \rightarrow Y) \leq \text{supp}(Y)(1-\alpha)$.

Se basant sur l'opérateur de fermeture φ à travers la relation d'équivalence canonique induite, ce résultat nous inspire à poser l'axiome d'inférence (AIR-NAD) ci-dessous :

Lemme 9 *: Axiome d'inférence sur les règles négatives approximatives (AIRNAD).*

Si $X \rightarrow \overline{Y}$ *et* Z, T *sont tels que* $\varphi(X) = \varphi(Z)$ *et* $\varphi(Y) = \varphi(T)$, *alors* $Z \rightarrow \overline{T}$.

La proposition suivante montre la correction de l'axiome (NA). Elle se démontre de façon analogue à la Proposition 22.

Proposition 23 . *L'axiome d'inférence* (NA) *est correct pour les règles négatives approximatives* $(\mathrm{M_{GK}}, \alpha)$*-valides, i.e., toute règle d'association déduite par application de (PA) à partir d'une règle négative approximative* $(\mathrm{M_{GK}}, \alpha)$*-valide est négative approximative* $(\mathrm{M_{GK}}, \alpha)$*-valide.*

Le théorème 6 ci-dessous caractérise la base que nous proposons pour les règles négatives approximatives $(\mathrm{M_{GK}}, \alpha)$-valides. Il se démontre de façon analogue au Théorème 5.

Thèorème 6 *[FDT06b, DFT06] L'ensemble* $BRNAD(\alpha)$ *défini par*

$$BRNAD(\alpha) = \{X \rightarrow \overline{Y} : \varphi(X) = X, \varphi(Y) = Y, 0 < \mathrm{conf}(X \rightarrow Y) \leq \mathrm{supp}(Y)(1-\alpha)\}$$

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

Exemple 4 *Considérons encore une fois le contexte présenté dans la Tableau 3. Aucune règle négative approximative n'est valide pour un seuil minimum de Support égal à* $\frac{2}{6}$ *et un seuil minimum de* $\mathrm{M_{GK}}$ *égal à* $\frac{2}{6}$*. Pour un seuil minimum de* $\mathrm{M_{GK}}$ *égal à* $\frac{1}{5}$*, on a* $BRNA(\frac{1}{5}) = \{BE \rightarrow \overline{AC}, AC \rightarrow \overline{BE}\}$.

La règle $BE \rightarrow \overline{AC}$ *est une règle de la base* $BNA(\frac{1}{5})$*. Par application de l'axiome d'inférence (NA), nous pouvons dériver les cinq règles d'association :*
$B \rightarrow \overline{A}, B \rightarrow \overline{AC}, E \rightarrow \overline{A}, E \rightarrow \overline{AC}, BE \rightarrow \overline{A}$.

Remarquons que si X et Y sont deux motifs tels que X et Y sont comparables, (*i.e.*, ou bien $X \subseteq Y$ ou bien $Y \subseteq X$), alors X favorise Y et réciproquement. Cela permet, pour un motif fermé X, de restreindre l'espace de recherche des motifs négatifs conséquents potentiels de X aux fermés incomparables avec X.

Un algorithme de génération de bases $BRNAD$ et $BRPA$ respectivement pour les règles négatives et positives approximatives est proposée dans la thèse de D.R. FENO [Fen07].

Au final, pour un seuil $\alpha \in]0, 1[$ fixé de $\mathrm{M_{GK}}$, la base $BRMGK(\alpha)$ des règles d'association $(\mathrm{M_{GK}}, \alpha)$-valide la réunion de ses cinq sous-bases : $BRPE$, $BRNE$, $BRPA(\alpha)$, $BRANG(\alpha)$, $BRAND(\alpha)$, soit : $BMGK(\alpha)$ $= BRPE \cup BRNE \cup BRPA(\alpha) \cup BRNAG(\alpha) \cup BRNAD(\alpha)$.

3.5 Conclusion partielle

Ces propriétés montrent que la mesure de qualité M_{GK} permet de sélectionner moins de règles que la mesure Confiance si l'on se borne aux règles positives, mesure conjointement l'écart à l'indépendance et le degré d'implication statistique entre deux motifs, respecte la condition d'équilibre pour un grand volume de données, et permet également d'avoir une vision unificatrice au sein d'un groupe majoritaire des mesures de qualité des règles d'association. Cette approche de normalisation apporte donc un nouvel éclairage sur l'ensemble des mesures de qualité. Une mesure normalisée a entre autres l'avantage d'élaguer systématiquement les règles dont prémisse et conséquent sont indépendants. Grâce à sa cohérence avec l'attraction et la répulsion entre deux motifs, elle est moins ambigüe et plus intelligible que le test de Khideux d'indéndance et la traditionnelle $Confiance$. En outre la mesure M_{GK} est favorablement plus discriminante que $Confiance$. Néanmoins, eu égard au caractère intuitif du mot *confiance* dans le langage social et au concept de probabilité conditionnelle qu'on lui fait incarner, il s'avère utile de garder ce terme, mais cette fois pour les règles M_{GK}-valides uniquement.

Chapitre 4

Extensions : Contextes quantitatif et complexe

4.1 Contexte quantitatif : règle d'association quantitative

4.1.1 Motivations

Depuis le début de la deuxième moitié des années 1990, des travaux s'intéressent aux attributs quantitatifs, par opposition à ceux qualitatifs ou booléens. En effet, constatant une faiblesse du fait de se contenter des données booléennes, sous le modèle du classique problème de paniers de clients dans un supermarché, R. Agrawal et al. [SA96] posent le problème ouvert suivant :" *We did not consider the quantities or values of the items bouth in a transaction which are important for some application. Finding such rules needs further work*".

Afin de donner un élément de réponses à ce questionnement si pertinent, nous présentons ici une méthode de l'extraction de règles d'association à partir des données quantitatives, notamment celles des paniers de clients en considérant, cette fois, la quantité effective de chaque article. Ainsi par exemple, pour m types d'articles $a_1, a_2, \ldots, a_m$ vendus, nous examinons les n transactions $(\alpha_{11}a_1, \alpha_{12}a_2, \ldots,$
$\alpha_{1m}a_m), \ldots, (\alpha_{n1}a_1, \alpha_{n2}a_2, \ldots, \alpha_{nm}a_m)$, où α_{ij} désigne la valeur de l'attribut (ou article) a_j existant dans la transaction de rang i.

Le problème de fouille des règles d'association à partir des données quantitatives a été introduit par [SA96] : l'algorithme qui y est proposé produi des règles d'association à partir d'une partition du domaine de chaque attribut quantitatif en intervalles et, ce, en transformant ainsi les données initialement quantitatives en données booléennes ou catégorielles. Plusieurs algorithmes ont été proposés pour traiter la fouille de règles d'association à partir des données catégorielles [STB+01].

Certes, des algorithmes courants de la fouille des règles d'association à partir des données quantitatives ont déjà permis d'appréhender utilement des variables quantitatives.

Cependant, de tels algorithmes introduisent quelques problèmes, entre autres, l'optimisation de la partition en intervalles, le caractère non intuitif et souvent non cohérent avec la perception humaine du choix de la partition, la non évidence de la distinction du degré d'adhésion pour la méthode d'intervalle.

La perte d'information croît lorsque le pas de discrétisation augmente [HGN00]. Ce qui a amené certains auteurs à proposer d'autres techniques d'extraction des règles d'association à partir de données quantitatives. Par exemple, pour contourner les sauts par constinuité produite par la discrétisation des attributs quantitatifs en intervalles disjoints deux à deux, certains travaux [Gey00] font une approche utilisant le concept d'ensemble flou : on fouille des règles d'association floues avec normalisation ; la considération d'ensembles flous *adoucit* les sauts de discontinuité évoqués ci-dessus.

Le nombre des règles d'association floues ainsi extraites demeure alors nettement inférieur à celui obtenu par la technique de discrétisation.

Un algorithme basé sur les notions de *support, moyenne et variance*, et utilisant deux tests de comparaison de moyenne et de variance par rapport à celles relatives à la totalité de la base de données, a été proposé par Aumann et Lindel (1999) [AL99] pour extraire des règles d'association quantitative de la forme :
âge$\in [70, 80] \Longrightarrow$(Poids moyen$= 90kg$). Cet algorithme se justifie par le fait que dans le cas particulier de variables binaires, la moyenne conditionnelle se réduit à la probabilité conditionnelle, soit à la notion habituelle de *confiance*. On a ainsi des règles exclusivement associatives, non interprétables en termes de pr'ediction malgré le caractère quantitatif des variables en jeu. La théorie est reprise et étendue dans [Web01] en définissant le concept de *règle d'impact (impact rule)*.

4.1.2 Approche proposée

Nous nous inspirons des techniques d'extraction des règles d'association sur des variables booléenes à base de motifs fréquents. nous partons de deux concepts statistiques le *coéfficient de dispersion (ou de variation)*, noté CD, et le *rapport de corrélation*, noté $\eta_{c/p}$, de la *variable conséquent* étant donnée la *variable premisse*, un test statistique validant l'intérêt de la règle d'association. Nous montrerons que l'inverse du coéfficient de dispersion généralise en quelque sorte la notion de fréquence aux variables quantitatives. D'autre part, en plus du concept classique de confiance, le niveau de confiance du test sur la signification du rapport de corrélation constituera une contrainte probabiliste supplémentaire rassurant sur la crédibilité de la règle d'association quantitative extraite. Sous la première condition que le rapport de

corrélation $\eta_{Y/X}$ soit significatif, les règles sont du type globale $X \Longrightarrow Y$, ou du type locale $(X \in Val(X)) \Longrightarrow (Y \in Val(Y))$, où $Val(X)$ désigne un ensemble de valeurs possibles de X, dont la qualité sera mesurée à l'aide de la mesure implicative orientée normalisée $\mathrm{M_{GK}}$.

Une règle globale $X \Longrightarrow Y$ peut s'interpréter, entre autres, par : X *influence* Y, il existe une dépendance fonctionnelle continue non linéaire ou linéaire entre X et Y ou une régression de Y en X, la connaissance de X permet de prédire quantitativement Y, etc.

Alors qu'une règle locale $(X \in Val(X)) \Longrightarrow (Y \in Val(Y))$ s'interprète soit de faon analogue à toute règle d'association booléenne classique, soit tout simplement en termes de dépendance fonctionnelle où Y prendrait telle valeur lorsque X prend telle valeur à une crédibilité statistique près. Cette approche permettrait de compenser la perte d'information due à la transformation d'une donnée quantitative en donnée booléenne ou catégorielle. Par ailleurs, les règles d'association quantitative ainsi extraites sont transitives ou quasi-transitives.

Lemme 10 *On perd généralement de l'information en réduisant deux variables effectivement à valeurs réelles positives ou nulles en deux variables binaires.*

En effet, soient X et Y deux variables prenant effectivement leurs valeurs réelles positives ou nulles dans respectivement $X(\Omega) = x_1, x_2, \ldots, x_r$ et $Y(\Omega) = y_1, y_2, \ldots, y_s$, $n_{i\bullet}$ et $n_{\bullet j}$ désignant les effectifs marginaux respectifs : alors la quantité d'information contenue dans le tableau de contingence correspondant est la somme des deux variances (qui est égale à la trace de la matrice de variances-covariances) : Soit la quantité d'information :

$$I_{v-reelle} = V(X)+V(Y) = \frac{1}{N}(\sum_{i=1}^{r} n_{i\bullet}x_i^2+\sum_{j=1}^{s} n_{\bullet j}y_j)^2)-\frac{1}{N^2}(\sum_{i=1}^{r} n_{i\bullet}x_i+\sum_{j=1}^{s} n_{\bullet j}y_j)^2$$

donc $N.I_{v-reelle}$ est de lÕordre de grandeur égal à $\sum_{i=1}^{r} n_{i\bullet}x_i^2+\sum_{j=1}^{s} n_{\bullet j}y_j)^2$, pour N assez grand, où $N = \sum_{i=1}^{r} n_{i\bullet} = \sum_{j=1}^{s} n_{\bullet j}$.
D'autre part, en réduisant ces variables X et Y en variables binaires, la quantité d'information mise en jeu devient :
$I_{v-binaire} = \frac{1}{N}(n_X + n_Y) - \frac{1}{N^2}(n_x^2 + n_y^2)$.

Donc $NI_{v-binaire}$ est de l'ordre de $(n_X + n_Y)$ pour N assez grand.
Or $(n_X + n_Y) < \sum_{i=1}^{r} n_{i\bullet}x_i^2 + \sum_{j=1}^{s} n_{\bullet j}y_j)^2$
Il en résulte l'inégalité : $I_{v-binaire} < I_{v-reelle}$.
Ce résultat motive la recherche d'une mesure de la qualité des règles d'association quantitative directement à partir des valeurs effectivement prises par elles.
Rappelons les résultats sur la décomposition d'une variance. Nous considérons ici deux variables aléatoires réelles X et Y définies sur un ensemble Ω(qui

sera naturellement fini dans le cas concret d'une base de données). Nous nous plaçons dans le cas où leurs variances respectives $V(X)$ et $V(Y)$ existent. Concernant l'étude de la dépendance de Y par rapport à X, par exemple, le théorème de la variance totale s'exprime par l'équation $V(Y) = E(V(Y/X))+ V(E(Y/X))$, où $E(Y/X)$ désigne la moyenne conditionnelle de Y sachant X, $V(E(Y/X))$ la variance expliquée par la régression de Y en X, et $E(V(Y/X))$ qui sera notée abusivement par $V(Y/X)$ la variance résiduelle de Y conditionnellement à X. Les deux inégalités suivantes résultent de la formule de la variance totale :
$V(E(Y/X)) \leq V(Y)$ et $V(Y/X) \leq V(Y)$
En théorie de l'estimation, l'inégalité $V(Y/X) \leq V(Y)$ signifie que le conditionnement de Y par X réduit la variance de Y, car pour toute valeur x de X , on a : $V(Y/X = x) \leq V(Y)$.

Définition 16 *Définitions :*

(i) On appelle rapport de corrélation de Y en X le nombre réel positif ou nul $\eta_{Y/X}$ défini par son carré : $\eta^2_{Y/X} = \frac{V(E(Y/X))}{V(Y)}$, c'est-à-dire variance intergroupes sur variance totale : c'est le pourcentage de la variance totale expliquée par l'éventuelle régression de Y en X.

(ii) On appelle le coéfficient de liberté de Y par rapport à X le nombre réel positif ou nul $L_{Y/X}$ défini par son carré : $L^2_{Y/X} = \frac{E(V(Y/X))}{V(Y)}$, ou variance intragroupe sur variance totale, qui exprime la proportion de la variance marginale de Y non expliquée par l'éventuelle régression de Y en X : cette notion de liberté relative est déjà utilisée par les physiciens [Fém90] ; [Fém03]

Ces deux concepts visiblement duaux ont les propriétés suivantes. On démontre que (cf. par exemple [Fém03], pages 110-115) :

(i) $0 \leq \eta_{Y/X} \leq 1$ et et $0 \leq L_{Y/X} \leq 1$.

(ii) $L^2_{Y/X} = 1 - \eta^2_{Y/X}$.

(iii) En général, on a : $\eta_{Y/X} \neq \eta_{X/Y}$ et $L_{Y/X} \neq L_{X/Y}$: il n' y a pas de symétrie dans les éventuelles indépendance ou dépendance.

(iv) $L_{Y/X} = 0$ (ou $\eta_{Y/X} = 1$) $\Longleftrightarrow$ (Il existe une dépendance fonctionnelle continue de Y relativement à X : Y est alors totalement dépendante de X, cette relation n'étant pas nécessairement linéaire).

(v) X et Y statistiquement indépendantes $\Longrightarrow L_{Y/X} = L_{X/Y} = 1$(ou $\eta_{X/Y} = \eta_{X/Y} = 0$).

(vi) $L_{Y/X} = 1$ (ou $\eta_{X/Y} = 0$) n'entrane pas nécessairement que Y soit indépendante de X : on dira que Y est totalement libre par rapport à X, la connaissance de X n'apporte aucune information sur Y.

(vii) Si $L_{X/Y} = 1$ (ou $\eta_{X/Y} = 0$), alors $\forall k \in \mathbb{N}, E(X^k Y) = E(X^k)E(Y)$, et donc en particulier $correlation(X, Y) = 0$, les deux variables ne sont pas linéairement corrélées.

(viii) Un rapport de corrélation est toujours supérieur ou égal à la valeur absolue du coéfficient de corrélation linéaire en général : $\eta_{X/Y} \geq |(correlation(X,Y))|$.
(ix) Si ($\eta_{X/Y} = |(correlation(X,Y))|$ ou ($\eta_{X/Y} \backsimeq |(correlation(X,Y))|$), alors il y a régression linéaire ou quasi-linéaire de Y en X .
(x) Si $Y = aX + b$, alors ($\eta_{X/Y} = |(correlation(X,Y))|$.

Les carrés $L^2_{Y/X}$ ou $\eta^2_{Y/X}$ peuvent ainsi mesurer le degré ou l'intensité d'interdépendance générale de Y en X de faon pas nécessairement symétrique, mais il s'avère préférable de travailler avec les écarts-types qui sont sémantiquement plus étroitement liés au caractère aléatoire des variables X et Y (ils sont exprimés par les mêmes unités que X et Y). Ainsi, le rapport de liberté de Y par rapport à X est défini par $L_{Y/X} = \sigma_{Y/X}/\sigma_Y$: il représente la moyenne du rapport de réduction de l'écart-type résiduel par rapport à l'écart-type total de Y. Donc, vu la propriété commune de non symétrie, contrairement au coéfficient de corrélation linéaire, ces deux rapports définis par $L_{Y/X}$ ou $\eta_{Y/X}$, dont les interprétations sont duales, apparaissent bien adaptés pour extraire une règle d'association avec dépendance fonctionnelle continue sur des variables quantitatives. La proposition (viii) signifie que généralement le coéfficient de corrélation linéaire sous-estime l'intensité de la liaison entre les deux variables, alors que les rapports de liberté ou de corrélation en donnent une idée plus exacte, à un test de signification près. Entre autres, une telle analyse conviendrait d'être faite avant même de passer à celle de régression, car elle permet d'identifier les variables dépendantes et celles indépendantes.

4.1.3 Le paradigme de l'approche proposée

(i) En général, le rapport de corrélation diffère du coéfficient de corrélation linéaire, sauf dans le cas où la distribution du couple (X,Y) est normale. Ce qui renforce l'intérêt du premier par rapport au second dans cette problématique de fouille des règles d'association.
(ii) Relation avec la mesure implicative de qualité normalisée $\mathrm{M_{GK}}$: Plaçons-nous dans le cas où les variables X et Y sont binaires : $X(\Omega) = Y(\Omega) = \{0;1\}$. L'espérance conditionnelle de Y sachant X prend les deux valeurs suivantes : $\overline{Y_0} = E(Y/X = 0) = 0.P(Y = 0/X = 0) + 1.P(Y = 1/X = 0) = \frac{n_Y - n_{XY}}{n - n_X}$, avec $n_{XY} = card(X' \cap Y')$. $\overline{Y_1} = E(Y/X = 1) = 0.P(Y = 0/X = 1) + 1.P(Y = 1/X = 1) = \frac{n_{XY}}{n_X} = P(Y/X) = conf(X \rightarrow Y)$. Par ailleurs, l'espérance et la variance totales de Y valent respectivement : $E(Y) = n_Y/n$, et $V(Y) = \frac{n_Y}{n}(1 - \frac{n_Y}{n})$. Par conséquent, le carré du rapport de corrélation de Y en X devient :

$$\eta_{Y/X} = \begin{cases} \sqrt{(\frac{P(X')}{1-P(X')}\frac{1-P(Y')}{P(Y')}\frac{(P(Y')-P(X)-P(X')P(Y'))}{P(Y')})}\mathrm{M_{GK}}^{f}(X\rightarrow Y), & \text{si } X \text{ fav. } Y \\ -\sqrt{\frac{P(Y')P(X')}{1-P(Y')}+\frac{(P(X')P(Y'))^2}{1-P(X')}}\mathrm{M_{GK}}^{d}(X\rightarrow Y), & \text{si } X \text{ défav. } Y. \end{cases} \quad (4.1)$$

Ainsi, à une faible valeur positive de α correspond un support f_X élevé ; et à une forte valeur de α correspondrait une faible valeur f_X du support de X. Ainsi le coéfficient de dispersion joue en quelque sorte le rôle du support utilisé dans le cas binaire, mais de façon inversement proportionnelle à son carré : cette fois les variables seront considérées selon le degré de dispersion croissant, soit de la plus homogène vers la moins homogène. Il doit donc figurer parmi les contraintes dans la fouille des règles d'association quantitative. Pour se fixer les idées, notons, par exemple, qu'à $\alpha = CV_{max} = 1$ correspondrait $f_X > 0,5$.

(iii) **À propos du coéfficient de dispersion :**
Rappelons que le coéfficient de dispersion d'une variable aléatoire réelle X est définie par $CD(X) = \sigma_X/|E(X)|$: cet indice permet de comparer deux dispersions. La restriction de sa considération à une variable binaire permet de comprendre son rapport avec la notion de support d'un attribut. En effet, supposons le cas de X binaire prenant ses valeurs dans la paire $\{0;1\}$, 1 désignant la présence de X dans une transaction considérée, par exemple. Alors sa moyenne serait tout simplement réduite à sa fréquence relative, c'est-à-dire à son support : Soit $E(X) = f_X = n_X/n = \mathrm{supp}(X)$; et son écart-type est égal à $\sigma_X = \sqrt{f_X(1-f_X)}$, donc $CD(X) = \sqrt{f_X(1-f_X)}/f_X$; par conséquent $CD(X) < \alpha \iff f_X > 1/(1+\alpha^2)$, ou encore $\frac{1}{CD(X)} > \alpha \iff f_X > \frac{\alpha^2}{1+\alpha^2}$.

Définition 17 ***Mesure de la qualité des règles globales d'association quantitative :***
On définit la mesure d'une règle globale $(X \Longrightarrow Y)$ *par la valeur du rapport de corrélation deY par rapport à* X *:* $mesure(X \Longrightarrow Y) = \eta_{Y/X} = \sqrt{1 - L^2_{Y/X}}$.
Soit à l'aide du concept de moyenne :
$\eta_{Y/X} = \sqrt{\frac{\sum i=1^r n_{i\bullet}(E(Y/X=x_i)-E(Y))^2}{\sum j=1^s n_{\bullet j}(y_j-E(Y))^2}}$

Dans la pratique, cette dernière expression se prête plus facilement au calcul. En général, l'implication, au sens de $\eta_{Y/X}$ significatif, va dans le sens croissant des coéfficients de dispersion de la prémisse vers le conséquent.

Proposition 24 *Les deux propriétés suivantes sont immédiates.*

(i) Les règles globales d'association quantitatives ainsi définies sont transitives
(i) ***Extension :*** *Elles se généralisent assez naturellement dans le cas où la variable prémisse est un vecteur aléatoire et le conséquent une variable aléatoire réelle.*

En effet : (i) La transitivité résulte naturellement du fait que la composition de deux applications est une application. Ainsi, la transitivité est évidente pour les règles d'association exactes ; elle continue de l'être, modulo une crédibilité près, pour les règles approximatives. (ii) Soit $X = (X_1, X_2, \ldots, X_n)$ un vecteur aléatoire défini sur le même ensemble que la variable aléatoire réelle Y. Les quantités conditionnelles $E(Y/X = x)$ et $V(Y/X = x)$, où $x = (x_1, x_2, \ldots, x_r)$ continuent d'avoir un sens. Notons que la décision sur la signification du rapport de corrélation de Y à X se prend selon la valeur de la statistique F de Fisher ou Z avec :
$F_{obs} = \frac{n-s}{s-1}\frac{\eta^2_{Y/X}}{(1-\eta^2_{Y/X})}$ et $Z_{obs} = \frac{1}{2}\ln(\frac{n-s}{s-1}\frac{\eta^2_{Y/X}}{(1-\eta^2_{Y/X})})$, les degrés de liberté de la loi de Fisher-Snedecor $F_{(\nu_1,\nu_2)}$ étant respectivement $\nu_1 = s-1$ et $\nu_2 = n-s$. Ce type de règle globale d'association quantitative vient naturellement compléter l'information contenue dans la règle d'association qualitative correspondante obtenue par la considération des variables binaires.

4.1.4 Algorithme d'extraction de règles globales d'association quantitative

Pour fixer les idées, examinons le cas de deux variables (aléatoires) réelles X et Y. Supposons que l'on ait $X(\Omega) = \{x_1, x_2, \ldots, x_r\}$ et $Y(\Omega) = \{y_1, y_2, \ldots, y_s\}$, avec les effectifs marginaux respectifs $n_{i\bullet}$ = nombre de fois où la valeur x_i de Xest prise dans Ω, pour $i$allant de 1 àr, puis $n_{\bullet j}$ =nombre de fois où la valeur y_j de Y est prise dans Ω, pour j allant de 1 à r. On a alors les étapes de calcul suivantes :

(i) Donner la valeur positive seuil CD_{max} des coéfficients de variations et le degré de crédibilité p des règles à extraire : Considérer les variables X et Y parmi celles dont les coéfficients de dispersion sont inférieures ou égaux au seuil fixé CD_{max}.

(ii) Croiser X et Y pour obtenir un tableau de contingence permettant d'avoir les diverses fréquences : fréquences jointes n_{ij}, marginales $n_{i\bullet}$ et $n_{\bullet j}$ respectivement de X et Y :

(iii) Calculer les moyennes conditionnelles de Y sachant les valeurs de X : Pour $i = 1$ jusqu'à $i = r$, faire $\overline{Y_i} = \frac{1}{n_{i\bullet}}\sum_{j=1}^{j=s} n_{ij}y_j$.

(iv) Calculer les variances conditionnelles de Y sachant les valeurs de X : Pour$i = 1$ jusqu'à $i = r$, faire $V(Y/X = x_i) = \frac{1}{n_{i\bullet}}\sum_{j=1}^{s} y_j^2 - (\overline{Y_i})^2$.

(v) Calculer la variance conditionnelle de Y sachant X : la moyenne des r variances conditionnelles : $V(Y/X)$, ainsi que celle de X sachant Y. Calculer de la même manière celle de X sachant Y : V(X/Y).

(vi)) Calculer les variances de X et de Y : V(X) et $V(Y)$.

(vii) Comparer les deux variances conditionnelles $V(Y/X)$ et $V(X/Y)$.

(viii) Calculer le coéfficient de corrélation linéaire $\rho(X, Y)$.

(ix) Calculer la liberté conditionnelle selon le résultat de la comparaison faite à l'étape (v). Si $V(Y/X) > V(X/Y)$, alors calculer $L_{Y/X}$

et $\eta_{Y/X}$, puis tester la signification de $\eta_{Y/X}$ et celle de la différence $\delta_{XY} = \eta_{Y/X} - |\rho|$ par rapport à zéro :
Décision : si $\eta_{Y/X}$ est significatif au seuil fixé p, alors retenir la règle globale $Y \Longrightarrow X$, avec le signification p, et si δ_{XY} est significativement différente de zéro, alors la liaison fonctionnelle continue de X en Y n'est pas linéaire, sinon cette liaison est linéaire (la fonction de régression étant croissante si $\rho > 0$, décroissante sinon,
sinon $\eta_{Y/X} \simeq 0,707$ et tester la crédibilité de $\eta_{Y/X}$: si $\eta_{Y/X}$ est significatif au seuil p, alors retenir l'équivalence statistique $X \Longleftrightarrow Y$ au seuil de signification égal à p. Notons que cet algorithme demeure valide dans le cas le plus général où la variable prémisse est multidimensionnelle et le conséquent une variable réelle, d'après la proposition 24.

4.1.5 Extraction des règles locales d'association quantitative

4.1.5.1 Objectif et position du problème

Revenons aux notations proposées dans le paragraphe ci-dessus. Nous nous proposons ici de trouver des intervalles I et J tels que, pour $\eta_{Y/X} > \eta_{X/Y}$, et $\eta_{Y/X}$ étant significatif, si l'on note $x = (X \in I)$ et $y = (Y \in J)$, alors la mesure $\mathrm{M_{GK}}(x \rightarrow y)$ est significative : on obtiendra ainsi des règles de type $(X \in I) \Longrightarrow (Y \in J)$, les intervalles I ou J pouvant être réduits aux singletons, analogues à celles définies par Y. Aumann et Y. Lindell [AL99], sans pour autant se baser sur des tests d'égalité de moyennes et d'égalité de variances. Finalement, on obtiendra ainsi au moins six types possibles (on laisse à l'utilisateur le choix du (ou des) type(s) qui l'intéresserait (intéresseraient) de règles locales d'association quantitative, à savoir :

(i) Les règles d'association quantitatives point par point :
$(X = x_i) \Longrightarrow (Y = y_j)$ qui correspond à $I = \{x_i\}$ et $J = \{y_j\}$

(ii)) Les règles d'association quantitatives entre intervalles :
$(X \in I) \Longrightarrow (Y \in J)$, I et J tous les deux non réduits à un singleton, I et J pouvant être des intervalles centrés respectivement aux valeurs moyennes $E(X)$ et $E(Y)$, c'est-à-dire
$I = [E(X) - k\sigma_X, E(X) + k\sigma_X]$, et $J = [E(Y) - k\sigma_Y, E(Y) + k\sigma_Y]$.

(iii) Les règles d'association quantitative de type point-intervalle :
$(X = x_i) \Longrightarrow (Y \in J)$, l'intervalle J étant non réduit à un singleton ;

(iv) Les règles d'association quantitative de type intervalleÐpoint :
$(X \in I) \Longrightarrow (Y = y_j)$, l'intervalle I seul étant non réduit à un singleton ;

(v) Les règles locales d'association quantitative valeur moyenne-intervalle :
$(X = E(X)) \Longrightarrow (Y \in J)$, telle que $\mathrm{M_{GK}}((X = E(X)) \Longrightarrow (Y \in J))$ soit maximale ;

(vi) Les règles locales dÕassociation quantitatives intervalle-valeur moyen-

ne $(X \in I) \Longrightarrow (Y = E(Y))$, telle que :
$\mathrm{M_{GK}}((X \in I) \Longrightarrow (Y = E(Y)))$ soit maximale.

4.1.5.2 Algorithme d'extraction des règles locales d'association quantitative

Pour chaque couple de variables (X, Y), l'algorithme se schématise comme suit.

Première étape : Calculer les rapports de corrélation $\eta_{Y/X}$ et $\eta_{X/Y}$, et le coéfficient de corrélation $\rho(X, Y) = correlation(X, Y)$.

Deuxième étape : Si $\eta_{Y/X}, \eta_{X/Y}$ et $\rho(X, Y)$ sont tous non significatifs, alors X et Y ne sont pas associées ; Sinon, si $\rho(X, Y)$ n'est pas significatif, alors 1. Si $\eta_{Y/X} > \eta_{X/Y}$, alors (X et Y son corrélées de faon non linéaires) : Pour tous i et j faire : Croiser $(X = x_i)$ et $(Y = y_j)$ $r_{ij} = \mathrm{M_{GK}}((X = x_i) \Longrightarrow (Y = y_j))$
Si r_{ij} est significatif alors retenir la règle $R_{ij} = ((X = x_i) \Longrightarrow (Y = y_j))$ Sinon, si $\eta_{Y/X} < \eta_{X/Y}$ alors échanger X et Y et revenir en 1. Si $\rho(X, Y)$ est significatif, alors :

2. Si $\eta_{Y/X} > \eta_{X/Y}$ et $\eta_{X/Y} \simeq \rho(X, Y)$ alors la régression de X en Y est linéaire, mais non celle de Y en X , aller en 1. Si $\eta_{Y/X} < \eta_{Y/X} \simeq \rho(X, Y)$, et alors échanger X et Y et revenir en 2.
Si $\eta_{Y/X} \simeq \eta_{Y/X} \simeq \rho(X, Y)$, alors X et Y sont corrélées linéairement.

4.1.6 Conclusion partielle

La présente étude montre la faisabilité d'une approche de la fouille des données quantitatives sans recourir à une discrétisation, ni au test de comparaison des distributions de moyennes ou de variances. Elle offre une mesure de la qualité d'une règle globale d'association quantitative permettant d'identifier pertinemment l'existence d'une liaison fonctionnelle continue non nécessairement linéaire entre certaines variables de la base de données, sans pour autant faire une analyse de régression effective, et généralisant en partie celle basée essentiellement sur le concept de support et de confiance appliquée dans le cas de données booléennes ou catégorielles : cette fois les concepts statistiques de base sont le coéfficient de dispersion indiquant le degré d'homogénéité, le rapport de corrélation dont la signification ou crédibilité est assurée par un test statistique de Fisher-Snedecor. Notons que ces concepts statistiques sont relativement faciles à manager sur le plan de calcul. Par ailleurs, la possibilité d'éclater empiriquement les règles globales d'association quantitative ainsi extraites, grâce à la combinaison avec la mesure probabiliste implicative normalisée $\mathrm{M_{GK}}$ qui est non symétrique et prenant en compte une dépendance statistique orientée, sous forme d'intéressantes règles locales d'association quantitative, crédibilise la présente approche.

Néanmoins, il est loisible de confronter cette manière de générer les règles locales à l'approche consistant à identifier les ensembles de valeurs à associer par l'utilisation de la transformation de Fourier discrète, afin d'atteindre une stabilité. Notre futur travail consistera également à la rendre opérationnelle afin de pouvoir l'éprouver sur des données quantitatives relevant des domaines différents comme les sciences médicales et agronomiques, par exemples. Néanmoins, se pose la question d'optimisation de la pertinence des règles locales d'association quantitative en combinant la présente approche avec celle basée sur un test de comparaison des moyennes ou variances, ou avec d'autres approches existantes ?

4.2 Traitement d'un contexte complexe

4.2.1 Position du problème

Les règles d'association ont été largement étudiées dans le cadre de données binaires. Comment étendre les résultats sur des résultats d'enquêtes menées sur un ensemble d'individus ou sur des descriptions structurées d'espèces végétales ou animales ? Nous montrons que si l'ensemble des observations peut être plongé dans un inf demi treillis alors il est possible de définir des mesures de qualité comme dans le cas binaire. Dans un premier temps, nous reformulons la définition des règles d'associationclassiques et les mesures de qualité dans le contexte des treillis afin de les étendre ensuite au traitement de questionnaires et de données structurées.

4.2.2 Une representation de données binaires

Soit $(U, B = \{\mathbf{V} < \mathbf{F}\})$ un attribut binaire défini sur un ensemble d'individus O. Les valeurs de verite sont structuees sous forme d'une chaine a deux elements telle que la valeur vraie $\mathbf{V}$ soit inferieure a la valeur fausse $\mathbf{F}$. On peut considérer l'attribut U comme une application $U \in B^O$. On considère le treillis complémenté, $\mathcal{M} = (B^O; <, \vee, \wedge, \mathbf{1}, \mathbf{0}, \text{-})$ qui est tel que :

- $U \leq V \iff \forall o \in O, U(o) \leq V(o)$,
- $W = U \wedge V \iff \forall o \in O, W(o) = U(o) \wedge V(o)$, cad $W(o) = \mathbf{V}$ si $U(o) = \mathbf{V}$ ou $V(o) = \mathbf{V}$
- $W = U \vee V \iff \forall o \in O, W(o) = U(o) \vee V(o)$, cad $W(o) = \mathbf{V}$ si $U(o) = \mathbf{V}$ et $V(o) = \mathbf{V}$
- Le plus petit élément $\mathbf{0}$ est la fonction telle que $\forall o \in O, \mathbf{0}(o) = \mathbf{V}$,
- $\overline{U} = \dashv U$, telle que $\overline{U}(o) = \mathbf{V}$ si $U(o) = \mathbf{F}$ et $\overline{U}(o) = \mathbf{F}$ dans le cas contraire.
- $\mathbf{0} = \overline{U} \wedge U$,

4.2.3 Règle d'association sur un questionnaire

4.2.3.1 Représentation des données

Un questionnaire est un ensemble de questions Q posé à un ensemble d'individus O. Une question est représentée par un attribut simple A_q dont le domaine $D_q = dom(A_q)$ est fini. On suppose que $|Q| = p$. Les réponses possibles ou modalités de la question A_q sont les valeurs du domaine D_q. Une stratégie possible pour rechercher des règles d'association sur un questionnaire est de se ramener à un tableau binaire en adoptant le codage dit en disjonctif complet pour chaque question et d'appliquer des algorithmes classiques de recherche de règles d'association. Cette approche qui associe un vecteur binaire à chaque modalité de réponse présente les inconvénients suivants :

- Les modalités d'une même question sont supposées indépendantes, or elles sont exclusives, on ne peut donner qu'une réponse à une question donnée. Rechercher des règles qui les combinent est inutile.
- Les non réponses ne sont pas prises en compte. Un interviewé peut ne pas répondre à une question pour diverses raisons, car il n'a pas compris la question ou elle n'a pas de sens compte tenu des réponses données à des questions précédentes. Le tableau est alors incomplet pour certains individus interviewés. Quelles valeurs binaires adopter pour ces questions non renseignées ?

Pour prendre en compte les non-réponses , on représente par

- "$*$ "une non réponse qui exprime que toutes les valeurs sont possibles
- Lorsque pour une question la réponse "valeur impossible" est envisageable, on introduira une modalité particulière "$\perp$" dans le domaine de cette question pour prendre en compte cette situation : $D_q = dom(A_q) \cup \perp$.

On munit le domaine discret de chaque attribut d'une structure de inf-demi-treillis en considérant comme plus petit élément $*$

$$\mathcal{T}_q = * \oplus D_q.$$

Remarque 12 *Il peut se produire que les valeurs du domaine d'un attribut soit déjà organisées de manière hiérarchique, qui est un inf-demi-treillis.*

L'ensemble des réponses possibles au questionnaire a une structure de inf-demi-treillis comme produit des inf-demi-treillis $\mathcal{T}_q$

$$\mathcal{T} = \pi_q \mathcal{T}_q$$

Soit δ la fonction qui associe à un individu $o \in O$ sa description c'est-à-dire les réponses aux questions

$$\delta : O \longrightarrow \mathcal{T}$$

ainsi

$$\delta(o) =< A_1 : w_1, \ldots, , A_q : w_q, \ldots, A_p : w_p >\in \mathcal{T}$$

ou w_q est la réponse donnée par l'individu o à la question A_q.

Par convention, une valeur inconnue sera omise, par exemple, la description

$$< sexe : *, CSP : "ouvrier", Region : "parisienne" >$$

sera notée plus simplement

$$< CSP : "ouvrier", Region : "parisienne" >$$

4.2.3.2 Règles d'association

À chaque élément $u \in \mathcal{T}$, on associe une fonction binaire, ou motif, $U = \beta(u) \in B^O$ telle que

$$U(o) = \mathbf{V} \iff u \leq \delta(o)$$

La relation $u \leq \delta(o)$ s'interprète comme u est "plus générale que" $\delta(o)$ et $U(o) = \mathbf{V}$ signifie que le motif U reconnaît l'observation o.

L'ensemble des motifs relatif à un questionnaire considéré est l'ensemble

$$\mathcal{M} = \{U = \beta(u) | u \in \mathcal{T}\}$$

Une règle d'association est un couple $(U, V) \in \mathcal{M}^2$ sur lequel il est possible de calculer des mesures de qualité probabilistes. Dans la suite, le motif associé à un élément $u \in \mathcal{T}$ est noté par le majuscule U correspondant à la lettre notant l'élément en question.

Précisons l'expression d'un motif. Remarquons que les éléments $< A_q : u_q >\in \mathcal{T}$ pour $q \in Q$ et $u_q \in dom(A_q)$ sont les éléments minimaux de $\mathcal{T}$ qui couvrent le plus petit élément $*$. Notons $u =< A_1 : u_1, \ldots, A_p : u_p >\in \mathcal{T}$, un individu $\delta(o) =< A_1 : w_1, \ldots, A_p : w_p >$, et $(A_q : u_q)$ le motif élémentaire tel que :

$$(A_q : u_q)(o) = \mathbf{V} \iff u_q \leq w_q.$$

Par définition $U(o) = \mathbf{V} \iff u \leq \delta(o) \iff u_q \leq w_q$ pour $q = 1, \ldots, p$, il est alors facile de voir que

$$U = (A_1 : u_1) \wedge \ldots \wedge (A_p : u_p)$$

ou $\wedge$ est le "et logique". On omettra de représenter dans une expression les termes $(A_q : *)$. Une règle d'association, un couple $(U, V) \in \mathcal{M}^2$ s'ecrit si $u =< A_1 : u_1, \ldots, A_p : u_p >$ et $v =< A_1 : v_1, \ldots, A_p : v_p >$

$$(A_1 : u_1) \wedge \ldots \wedge (A_p : u_p) \longrightarrow (A_1 : v_1) \wedge \ldots \wedge (A_p : v_p)$$

4.2.4 Correspondance de Galois sur contexte complexe

On peut définir une correspondance de Galois entre l'inf-demi-treillis $\mathcal{T}$ et $\mathcal{P}(O)$ à travers les applications extension $ext : \mathcal{T} \longrightarrow \mathcal{P}(O)$ et intension $int : \mathcal{P}(O) \longrightarrow \mathcal{T}$ qui sont définis comme suit :

$$ext(u) = \{o \in O | u \leq \delta(o)\} = U^{-1}(\mathbf{V})$$

l'extension de u est l'ensemble d'observations reconnu par le motif U qui lui est associé. Si $L \subset O$

$$int(L) = \bigwedge_{o \in L} \delta(o)$$

c'est le plus petit motif qui reconnaît tous les observations de L, ou encore c'est le plus grand ensemble des réponses communes aux observations de L. On note $\mathcal{T}(O)$ l'inf-demi-treillis engendré par $\delta(O) = \{\delta(o) | o \in O\} \subset \mathcal{T}$, c'est-à-dire le plus petit inf-demi-treillis contenant $\delta(O)$.

4.2.5 Représentation des données arborescentes

On suppose que l'on dispose d'un ensemble d'attributs simples (A_j, D_j) d'identificateur A_j et de domaine D_j discret ou numérique. On appelle attribut structuré ou tuple une séquence

$$A =< A_1, \ldots, A_q, , \ldots, A_p >$$

où A_q est un attribut simple ou structuré. Le domaine cad l'ensemble D des valeurs d'un attribut structuré A est défini de manière inductive :

- Si A_q est un attribut simple $A_q : w_q$ est une valeur simple pour $w_q \in D_q$ et $A_q : w_q$ appartient à D
- Si $A_1, \ldots, A_p$ sont des attributs simples ou structurés, et $A_1 : w_1 \in D, \ldots, A_p : w_p \in D$ alors $A :< A_1 : w_1, \ldots, A_p : w_p >\in D$ est une valeur structurée.

Le modèle descriptif est défini par la donnée d'un attribut tuple $A =< A_1, \ldots, A_p >$. La description des individus $o \in O$ est donnée par une fonction

$$\delta : O \longrightarrow D$$

telle que

$$\delta(o) =< A_1 : w_1, \ldots, A_q : w_q, \ldots, A_p : w_p >\in D$$

avec : $w_q \in D_q$ ou D_q est le domaine de l'attribut A_q. C'est un arbre dont les noeuds sont les identificateurs A_q des attributs et les feuilles les valeurs des attributs simples.

4.2.6 Squelette

Les données sont représentées par des arbres décrits par un attribut tuple, mais dont des parties peuvent être absentes. Nous allons définir le *squelette* associé à une valeur structurée qui renseigne sur la présence ou absence de sous-arbre. On représente, comme précédemment, par

- $*$: un noeud qui n'est pas défini (valeur inconnue) mais possible.
- $\perp$: le noeud ne peut pas exister à cause de certaines valeurs des attributs pères (valeur impossible).

Définition 18 *Soient* $S = \{+, *, \perp\}$*, une valeur structurée* $d =< A_1 : w_1, \ldots, A_q : w_q, \ldots, , A_p : w_p >$ *tel que* $w_q \in D_q$*, pour* $q \in Q$*. Le squelette associé à* d *est l'arbre défini par la fonction* $\sigma : \bigcup_q D_q \longrightarrow S$ *comme suit*

- $\sigma(*) = *$
- $\sigma(\perp) = \perp$
- *Si* $w_q \notin \{*, \perp\}$ *alors* $\sigma(w_q) = +$

Il est noté

$$d_\sigma =< A_1 : \sigma(w_1), \ldots, A_q : \sigma(w_q), \ldots, A_p : \sigma(w_p) >$$

4.3 Classification

Dans cette dernière partie, nous proposons une approche, fondée sur la classification, pour rechercher des règles d'association dans un contexte de descriptions ordonnées, c'est-à-dire un contexte où l'espace de descriptions des objets est un ensemble (partiellement) ordonné. L'idée est de classifier l'ensemble des objets, puis de considérer comme candidat prémisse ou conséquent d'une règle d'asociation l'intension d'une des classes résultant de cette classification. Ainsi, un espace de recherche de règles d'association valides est entièrement déterminé par les classes obtenues, et variera selon la mesure de dissimilarité utilisée, la méthode de classification adoptée, ou la structure de classification construite. L'association des règles à des classes optimisant un critère est un facteur de pertinence qui renforcerait la qualité de ces règles, évaluée par ailleurs en utilisant une ou plusieurs des mesures de qualité proposées dans la littérature.

4.3.1 Le paradigme des règles d'association dans un contexte de descriptions ordonnées

Toutesfois, malgré ces contraintes, le nombre de règles d'association conformes à la définition ci-dessus reste très élevé. Ainsi, dans un soucis d'informativité et de pertinence, on n'en retient que ceux qui sont valides au sens d'une (un ensemble de) mesure(s) de qualité. La plupart de ces mesures de qualité sont probabilistes (*i.e.* se définissent entièrement à partir

d'un tableau de contingence) et les plus connues d'entre elles sont le support, et la confiance.

En fait, unn contexte binaire peut être vu comme un triplet $\mathbb{K} = (\mathcal{O}, \mathcal{P}(\mathcal{A}), \delta)$ où $\mathcal{P}(\mathcal{A})$ est l'ensemble ordonné $(\mathcal{P}(\mathcal{A}), \subseteq)$ et δ l'application qui associe à chaque objet o le motif constitué des items que possède l'objet o. Ainsi, pour chaque motif X, l'extension X' de X sera définie par $X' = \{o \in \mathcal{O} : X \subseteq \delta(o)\}$.

Cette représentation des contextes binaires permet de considérer naturellement les règles d'association dans un *contexte de descriptions ordonnées*, *i.e.*, un contexte $\mathbb{K} = (\mathcal{O}, \mathcal{D}, \delta)$ où $\mathcal{D} := (\mathcal{D}, \leq)$ est un ensemble ordonné et δ est une application qui associe à chaque objet o sa description $\delta(o)$ dans $\mathcal{D}$. L'ensemble $\mathcal{D}$ est alors appelé l'espace de description des objets. Les motifs d'un tel contexte seront des éléments de $\mathcal{D}$ et l'extension X' d'un motif $X \in \mathcal{D}$ sera définie par $X' = \{o \in \mathcal{O} : X \leq \delta(o)\}$ comme cela est défini pour des objets symboliques [Did95]. sectionRecherche de règles d'association

Soit $\mathbb{K} = (\mathcal{O}, \mathcal{D}, \delta)$ un contexte de descriptions ordonnées tel que la borne inférieure de deux descriptions $\delta(o_1)$ et $\delta(o_2)$ existe toujours dans $\mathcal{D}$. L'idée est de spécifier un espace de recherche de règles d'association dans $\mathbb{K}$ en classifiant l'ensemble $\mathcal{O}$ des objets du contexte. Plus précisément, il s'agit d'associer à un système $\mathcal{C}$ de classes dans $\mathcal{O}$, l'ensemble $R_{\mathcal{C}}$ de candidats prémisses ou conséquents de règles, défini par

$$R_{\mathcal{C}} = \{\text{int}(C) : C \in \mathcal{C}\}$$

où $\text{int}(C) = \inf\{\delta(x) \in \mathcal{D} : x \in C\}$ est l'intension de C. La qualité des règles d'association formées de couples d'éléments de $R_{\mathcal{C}}$ est alors évaluée en utilisant une ou plusieurs des diverses mesures de qualité de règles proposées dans la littérature.

Lorsque les données sont binaires et de grande taille, les difficultés à résoudre dans la recherche des règles significatives selon une ou plusieurs mesures de qualité probabilistes sont le grand nombre de règles potentielles et le temps important de lecture des données pour le calcul des mesures de qualité. La stratégie que nous proposons pour surmonter ces problèmes est de s'appuyer sur des classifications des objets pour dériver un ensemble de règles d'association significatives. Nous supposons que l'ensemble des objets est muni d'une dissimilarité d qui tient compte de l'ordre dont est muni l'espace de description et que celui-ci est stable par la borne inférieure. Si X' est un ensemble d'objets, on définit l'intension de X' comme $int(X') = \wedge_{o \in X'} \delta(o) \in \mathcal{D}$. L'idée est de classifier l'ensemble des objets par rapport à cette mesure de dissimilarité. Chaque classe de cette classification ainsi obtenue donnera lieu à un motif, son intension. L'ensemble de ces intensions correspondra à l'espace de recherche des règles d'association.

4.3.2 Exemple

Exemple 5 *Les attributs CSP, Dépense, Satisfaction, qui caractérisent les objets du tableau de données (Table 4.1) sont qualitatives. Pour de tels attributs* (A_q, D_q)*, on munit chaque domaine d'une structure de inf demi treillis en considérant* D_q *comme une antichaîne et* $*$ *comme le plus petit élément de* D_q*,* $(D_q; \leq, \wedge, *)$*. Une distance habituellement choisie pour les objets quand les attributs sont qualitatifs est celle du chi-deux. Lorsque le type d'un attribut* (A_q, D_q) *est multivalué, comme Hébergement, l'inf demi treillis considéré est* $(\mathcal{P}(D_q), \subseteq, \cap, \emptyset)$*. La distance différence symétrique ou celle proposée dans [Dia03] sont de bonnes distances candidates pour des objets quand l'attribut est multivalué. Pour les attributs de type intervalle réel, comme Durée, le domaine est un ensemble de valeurs de la forme* $u = [\underline{u}, \overline{u}]$ *avec* $\underline{u}, \overline{u}$ *réels tels que* $\underline{u} \leq \overline{u}$*. La relation d'ordre choisie est celle duale de l'inclusion. Elle exprime l'idée que plus un intervalle est large plus l'information qu'elle représente est imprécise. L'opérateur inf est* $u \wedge v = [\underline{u} \wedge \underline{v}, \overline{u} \vee \overline{v}]$*. Il est possible de définir plusieurs types de distance sur des objets caractérisés par un attribut de type intervalle [GD92], [Dia03].*

Soit un ensemble d'objets O décrits par des attributs dont chaque domaine est un inf demi treillis $\mathcal{T}_q = (D_q; \leq, \wedge, *)$ muni d'une dissimilarité locale d_q pour $q \in Q$. L'ensemble produit $\mathcal{T} = \Pi_q \mathcal{T}_q$ est un inf demi treillis comme produit des inf demi treillis $\mathcal{T}_q$ et $d = \Sigma_q d_q$ est une dissimilarité sur $\mathcal{T}$. L'espace de description des objets que nous condidérerons est l'inf demi treillis engendré par les descriptions des objets du tableau des données $\mathcal{D} = \{\wedge_{o \in O} \delta(o)\} \subset \mathcal{T}$.

A titre d'exemple, voyons comment une règle est déduite à partir de deux classes du tableau de données (Table 4.1).

Soient les classes $X'_1 = \{o_1, o_2, o_3, o_6, o_7\}$, et $X'_2 = \{o_1, o_2, o_3, o_4, o_5\}$. On a :
$int(X'_1) = (CSP = "ouvrier") \wedge (Hebergement = *) \wedge (Depense = *) \wedge (Satisfaction = *) \wedge (Duree = [3, 14] = *)$
que l'on réécrit plus simplement $int(X'_1) = (CSP = "ouvrier")$ car dans le calcul de l'extension les attributs pour lesquels les valeurs sont égales à $*$ le plus petit élément peuvent être omis. L'intension de la classe X'_2 est :
$int(X'_2) = (CSP = *) \wedge (Hebergement = "Gite, Camping") \wedge (Depense = *) \wedge (Satisfaction = *) \wedge (Duree = [3, 14] = *)$ ou $int(X'_2) = (Hebergement = "Gite, Camping")$.
On en déduit la règle :
r_1 : CSP="ouvrier" $\rightarrow$ Hébergement="Gite, Camping" avec $\chi^2(r_1) = 0.4$, $M_{GK}(r_1) = 0.2$, la dépendance est positive.

TAB. 4.1 – Le tableau de données "Tourisme"

	CSP	Hébergement	Dépense (Euros)	Satisfaction	Durée séjour
o_1	Ouvrier	Gite,Camping, Famille	5000-10000	Peu	[7,14]
o_2	Ouvrier	Gite,Camping	5000-10000	Oui	[7,10]
o_3	Ouvrier	Gite,Camping	5000-10000	Oui	[7,14]
o_4	Employé	Gite,Camping, Hotel	5000-10000	Oui	[3, 5]
o_5	Cadre	Gite,Camping,Hotel	10000-15000	Peu	[3, 6]
o_6	Ouvrier	Hotel	10000-15000	Peu	[7,14]
o_7	Ouvrier	Famille	2000-5000	Peu	[7,7]
o_8	Autres	Hotel	15000-20000	Peu	[7,14]
o_9	Autre	Famille	2000-5000	Non	[3,4]
o_{10}	Cadre	Gite	2000-5000	Non	[3,3]

Chapitre 5

Conclusion générale et Perspectives

5.1 Conclusion

L'approche formelle des mesures probabilistes de qualité normalisées des règles d'associations, partant des contextes binaires de fouille de données et étendue aux contextes quantitatifs et aux données complexes comprenant treillis et intervalles, prenant en compte les cinq situations de référence communément intuitives (indépendance, incompatibilité, implication logique, attraction et répulsion partielles), apporte un nouvel éclairage sur l'ensemble desdites mesures d'intérêt des règles d'association.

Ces critères objectifs d'évaluations des règles d'association se répartissent en quatre catégories : les mesures normalisées (continues, discontinues), les mesures dont la normalisée associée par une homéomorphie affine est la mesure de Guillaume-Kenchaff M_{GK}, les mesures dont la normalisée associée par une homéomorphie affine est différente de M_{GK}, et celles qui n'admettent pas de normalisée associée par une homéomorphie affine.

Parmi les mesures normalisées celles qui sont continuent offrent un réel intérêt dans une tâche d'extraction des règles d'association pertinentes.

Leur ensemble étant stable par combinaison linéaire convexe (ou addition barycentrique), il existe ainsi une infinité de mesures probabilistes de qualité de règles normalisées continues. Cependant, celles-ci s'expriment toutes en fonction de M_{GK} à coéfficients dynamiques près et à des puissances des composantes favorable et défavorable près.

La mesure de qualité M_{GK} joue un rôle important parmi les normalisées continues.

Elle apparaît aussi plus discriminante que l'indice pionnier confiance et a l'avantage partagé par toutes les mesures normalisées continues de ne pas produire des règles incohérentes, d'avoir une sémantique d'implication statistique, et d'être bien adaptée à générer des règles de quatre types : règles

postives exactes et approximatives, négatives(à droite ou bien à gauche) exactes et approximatives.

Il est possible d'en générer une base selon la mesure M_{GK}, avec un seuil de signification statistique à l'appui.

5.2 Perspectives

Les actions mentionnées ci-après s'inscrivent toutes dans la continuation des travaux présentés dans ce mémoire sur le plan mathématique (M) ou informatique (I).

(I) Implémentation et optimisation des algorithmes proposés dans la thèse de D. Feno pour la génération de base des règles d'association valides selon M_{GK} et $confiance$; génération de toutes les règles et comparaison expérimentale avec les différentes bases de la littérature.

(I) Élaboration d'un outil de fouille de données permettant d'expérimenter dans divers domaines : marketing, détection de fraude, épidémiologie, didactique de disciplines, sociologie, psychologie, traitement d'enquêtes, sciences physiques, etc.

(M) Recherche d'exemples d'homéomorphie rationnelle permettant de normaliser les mesures probabilistes de qualité qui n'admettent pas de normalisée associée par une homéomorphie affine.

(M) Étude topologique de l'ensemble $\mathcal{C}(\mathcal{N})$ vis à vis de l'ensemble des mesures de qualité normalisées discontinues, à l'instar de la densité de $\mathbb{Q}$ dans $\mathbb{R}$.

(M & I) Optimisation de la recherche d'une mesure de qualité normalisée continue la plus sélective (en termes de puissance et de seuil de signification) à partir d'une mesure normalisée continue initiale, par des jeux de puissances.

(M) Étude de la possibilité d'extension de l'indice de qualité M_{GK} en probabilités.

(M & I) Recherche d'un découpage optimale d'une variable quantitative X telle que $M_{GK}(X \to Y$ soit maximal, avec Y une variable qualitative.

(I) Extension de M_{GK} en fouille de données sur le web (web mining).

Bibliographie

[AIS93] Agrawal R., Imielinski T. et Swami A. Mining association rules between sets of items in large databases. *In : Proc. of the ACM SIGMOD International Conference on Management of Data*, éd. par Buneman P. et Jajodia S., pp. 207–216, 1993. Washington,U.S.A.

[AL99] Aumann Y. et Lindell Y. A statisticall theory for quantitative association rules. *In : Proc. KDD'99*, pp. 261–270, 1999. San Diego, CA.

[ALS03] Azé J., Lucas N. et Sebag M. Fouille de données visuelle et analyse de facteurs de risque médical. *In : EGC*, pp. 183–188, 2003.

[Arm74] Armstrong W. W. Dependency structures of data base relationships. *Information Processing*, vol. 74, 1974, pp. 580–583.

[AS94] Agrawal R. et Srikant R. Fast algorithms for mining association rules. *In : Proc. of the 20th VLDB Conference*, pp. 487–499, 1994. San Diego,Chile.

[AZ04] Antonie M. L. et Zaïane O. R. Mining positive and negative association rules : An approch for confined rules. *In : Proc. 8th Int. Conf. on Principle and Practice of Knowledge Discovery in Databases (PKDD'04)*. pp. 27–38, 2004. Springer-Verlag.

[Bay98] Bayardo R. J. Efficiently mining long patterns from databases. *In : Proc. of the ACM SIGMOD Conference*, pp. 85–93, June 1998. Washington,U.S.A.

[BGBG05] Blanchard J., Guillet F., Briand H. et Gras R. Assessing rule with a probabilistic measure of deviation from equilbrium. *In : Proc. of 11th International Symposium on Applied Stochastic Models and Data Analysis ASMDA*. pp. 191–200, 2005. Brest,France.

[BM70] Barbut M. et Monjardet B. *Ordre et classification*. Paris, Hachette, 1970.

[BMS97] Brin S., Motwani R. et Silverstein C. Beyond market baskets: Generalizing association rules to correlation. *In : Proc. of the ACM SIGMOD Conference*, pp. 265–276, 1997. Tucson, Arizona.

[BMUT97] Brin S., Motwani R., Ullman J. D. et Tsur S. Dynamic itemset counting and implication rules for market basket data. *In : Proc. of the ACM SIGMOD Conference*, pp. 255–264, 1997.

[Bor86] Bordat J. P. Calcul pratique du treillis de Galois d'une correspondance. *Math. Sci. Humaines*, vol. 96, 1986, pp. 31–47.

[Bri04] Brisson L. Mesure d'intérêt subjectif et représentation des connaissances, 2004. Technical Report ISRN I3S/ RR-2004-35FR, Université de Nice.

[BTP+02] Bastide Y., Taouil R., Pasquier N., Stumme G. et Lakhal L. Pascal : un algorithme d'extraction des motifs frquents. *Technique et science informatique*, vol. 21, 2002, pp. 65–95.

[BW95] Baker K. A. et Wille R. (édité par). *Lattice theory and its applications*. Berlin, Heldermann-Verlag, 1995.

[Cas99] Caspard N. A characterization theorem for the canonical basis of a closure operator. *Order*, vol. 16, 1999, pp. 227–230.

[CH90] Church K. W. et Hanks P. Word association norms, mutual information and lexicography. *Computational Linguistics*, vol. 16, 1990, pp. 22–29.

[CM00] Chaudron L. et Maille N. Generalized formal concept analysis. *In : Proc. 8th Int. Conf. on Conceptual Structures, ICCS'2000*, éd. par Mineau G. et Ganter B., pp. 357–370, 2000.

[CM03] Caspard N. et Monjardet B. The lattices of closure systems, closure operators, and implicational systems on a finite set: a survey. *Discrete Applied Mathematics*, vol. 127, 2003, pp. 241–269.

[Coh60] Cohen J. A coefficient of agreement for nominal scale. *Educational and Psychological Measurement*, vol. 20, 1960, pp. 37–46.

[CR93] Carpineto C. et Romano G. Galois: an order theoretic approach to conceptual clustering. *In : Proceedings of the Machine Learning Conference*, pp. 33–40, 1993.

[Day92] Day A. The lattice theory of functional dependencies and normal decompositions. *Internat. J. Algebra Comput.*, vol. 2, 1992, pp. 409–431.

[DdRFT07] Diatta Jean, danial R. Feno et Totohasina André. Mining bases for m_{GK}-valid association rules. *The Global Journal of Pure and Applied Mathematics, GJPAM, Research India Publications, (9 pages)*, vol. accepted to appear, 2007.

[DFT06] Diatta J., Feno D. R. et Totohasina A. Galois lattices and based for m_{GK}-valid association rules. *In : Proc. of the Fourth International Conference on Concept Lattices and Their Applications, CLA'06*, pp. 127–138, 2006. Hammamet, Tunisie.

[Dia03] Diatta J. A mixed measure of content on the set of real numbers. *Journal of Computational and Applied Mathematics*, vol. 151, 2003, pp. 85–105.

[Dia05] Diatta J. Caractérisation des ensembles critiques d'une famille de moore finie. *In : Douzièmes journées de la Société Francophone de Classification*, pp. 126–129, 2005. Montreal,Canada.

[Did95] Diday E. Probabilistic, possibilist and belief objects for knowledge analysis. *Annals of Operations Research*, vol. 55, 1995, pp. 227–276.

[DL04] Domenach F. et Leclerc B. Closure systems, implicational systems, overhanging relations and the case of hierarchical classification. *Mathematical Social Sciences*, vol. 47, 2004, pp. 349–366.

[DM00] Das G. et Mannila H. Context-based similarity measures for categorical databases. *In : Principles of Data Mining and Knowledge Discovery*, pp. 201–210, 2000.

[Dom02] Domenach F. *Structures latticielles, correspondances de* G*alois contraintes et classification symbolique*. France, Thèse de PhD, Université Paris 1 Panthéon-Sorbone, 2002.

[DP94] Davey B. A. et Priestley H. A. *Introduction to Lattices and Orders*. Cambridge, Cambridge University Press, 4th edition, 1994.

[DRT07] Diatta J., Ralambondrainy H. et Totohasina A. Towards a unifying probabilistic implicativenormalized quality measure for association rules. *Quality Measures in Data Mining*, 2007, pp. 237–250.

[EP96] Elder J. et Pergibon D. A statistical perspective on knowledge discovery in databases. *AAAI Press*, 1996, pp. 83–115.

[Eve55] Everett C. J. Closure operators and Galois Theory in Lattices. *Trans. Amer. Math. Soc.*, vol. 55, 1955, pp. 514–525.

[FDT06a] Feno D. R., Diatta J. et Totohasina A. Normalisée d'une mesure probabiliste de qualité des règles d'association : étude des cas. *In : Actes du 2nd Qualité des Données et des Connaissances (D.K.Q.)*, pp. 25–30, 2006. Lille, France.

[FDT06b] Feno D. R., Diatta J. et Totohasina A. Une base pour les règles d'association d'un contexte binaire valides au sens de la mesure de qualité m_{GK}. *In : Proc. of the 13ème Rencontre de la Société Francophone de Classification*, pp. 105–109, 2006. Metz, France.

[FDT07] Feno D. R., Diatta J. et Totohasina A. Génération de bases pour les règles d'association m_{GK}-valides. *In : Proc. of the 14ème Rencontre de la Société Francophone de Classification*, pp. 101–104, 2007. Paris, France.

[Fen07] Feno D.R. *Mesures de qualité des règles d'association : normalisation et caractérisation de bases*. France, Thèse de PhD, Université de La Réunion, 2007.

[Fer06] Ferré S. *Systèmes d'information logique : un paradigme logico-contextuel pour interroger, naviguer et apprendre*. France, Thèse de PhD, Université de Rennes I, 2006.

[Fém90] Féménias J.L. Interdependence of parameters in spectroscopic data reduction. *Journal of Mollecular Spectroscopy*, vol. 144, 1990, pp. 212–223.

[Fém03] Féménias J.L. *Probabilités et statistiques pour les sciences physiques*. Paris, Dunod, 2003.

[FPSM92] Frawley W. J., Piatetsky-Shapiro G. et Matheus C. J. Knowledge discovery in databases - an overview. *Ai Magazine*, vol. 13, 1992, pp. 57–70.

[FPSS96] Fayyad U. M., Piatetsky-Shapiro G. et Smyth P. Knowledge discovery and data mining : towards a unifying framework. *In : Proceedings of the second International Conference on Knowledge Discovery and Data Mining*, pp. 82–88, 1996. Portland, OR.

[Fre99] Freitas A. On rule interestingness measures. *Knowledge-Based System*, vol. 12, 1999, pp. 309–315.

[GD86] Guigues J. L. et Duquenne V. Famille non redondante d'implications informatives résultant d'un tableau de donnes binaires. *Mathmatiques et Sciences humaines*, vol. 95, 1986, pp. 5–18.

[GD92] Gowda K.C. et Diday E. Symbolic clustering using a new similarity measure. *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, 1992, pp. 368–378.

[GDGB07] Gras R., David J., Guillet F. et Briand07 H. Stabilié en a.s.i. de l'intensité d'implication et comparaisons avec d'autres indices de qualité de règles. *In : Actes du 3e Atelier Qualité des Données et des Connaissances, EGC'07*, pp. 35–43, 2007.

[Gey00] Geynessei A. *A fuzzy approach for mining association rules*. Rapport technique 336, Turku Center For Computer Science, 2000.

[GH06] Geng L. et Hamilton H. J. Interestingness measures for data mining: A Survey. *ACM Computing Surveys*, vol. 38, 2006, pp. 1–31.

[GKCG01] Gras R., Kuntz P., Couturier R. et Guillet F. Une version entropique de l'intensité d'implication pour les corpus volumineux. *Extraction des Connaissances et Apprentissage*, vol. 1, 2001, pp. 69–80.

[Goo65] Good I.J. The estimation of probabilities: An essay on modern bayesian methods. *The MIT press*, vol. MA, 1965.

[GS88] Goodman R. et Smyth P. Information theoretic rule induction. *In : Proc. of the ECAI-88*, pp. 357–362, 1988. Munich,Germany.

[GSB+96] Gras R., S. Almouloud, Bailleul M., Larher A., Polo M., Ratsimba-Rajohn H. et Totohasina A. *L'implication statistique. Nouvelle méthode exploratoire de données*. Grenoble, France, La Penée sauvage, 1996.

[GT95a] Gras R. et Totohasina A. Chronologie et causalité, sources d'obstacles épistémologiques à l'apprentissage de probabilité conditionnelle. *Recherche en Didactique des mathématiques*, vol. 15, 1995, pp. 49–95.

[GT95b] Gras R. et Totohasina A. Conceptions d'élève sur sur la notion de probabilité conditionnelle relevées par une méthode d'analyse de données : implication-similarité-corrélation. *Educational studies in mathematics. Kluwer academic publisher*, vol. 28, 1995, p. 21 pages.

[Gu90] Gunoche A. Construction du treillis de Galois d'une relation binaire. *Mathématiques Informatique et Sciences humaines*, vol. 109, 1990, pp. 41–53.

[Gui00] Guillaume S. *Traitement des données volumineuses. Mesures et algorithmes d'extraction des règles d'association et règles ordinales*. France, Thèse de PhD, Universté de Nantes, 2000.

[GW99] Ganter B. et Wille R. *Formal concept analysis, Mathematical foundations*. Berlin, Springer Verlag, 1999.

[HGB05a] Huynh X., Guillet F. et Briand H. Une plateforme exploratoire pour la qualité des règles d'association : Apport pour l'analyse implicative. *In : Proc. of Troisièmes Rencontres Internationales A.S.I.*, pp. 339–349, 2005. Palermo,Italie.

[HGB05b] Huynh X. H., Guillet F. et Briand H. Arqat : An exploratory analysis tool for interstingness measures. *In : Proc. of Applied Stochastic Models and Data Analysis*, pp. 334–344, 2005. Brest,France.

[HGN00] Hipp J., Güntzer U. et Nakhaeizadeh G. Algorithms for Association Rule Mining - a General Survey and Comparison. *SIGKDD Explorations*, vol. 2, 2000, pp. 58–64.

[HH99] Hilderman R. J. et Hamilton H. J. Knowledge discovery and interestingness measures: A survey, 1999. Technical Report CS 99-04, Department of Computer Science, University of Regina.

[HYS05] Hamrouni T., Yahia S. Ben et Slimani Y. Prince : An algorithm for generating rule bases without closure computations. *In : Proc. of the 7th DaWaK Conference*, pp. 346–355, 2005.

[Ise51] Iseki K. On closure operation in lattice theory. *In : Idang. Math 13*, éd. par 54 Nerd. Akad. Wetensch Proc. Ser. A, pp. 318–320, 1951.

[JHA07] J.Diatta, H.Ralambondrainy et A.Totohasina. Règles d'association dans un contexte de descriptions ordonnées. *In : XIVmes Rencontres SFC 07, Socié Franaise de Classification, 5-7 sept,*, éd. par 07 SFC, 2007. TELECOM Paris.

[KM06] Kurgan L.A. et Musilek P. A survey of knowledge discovery and data mining process models. *The knowledfge Engineering review, U.K.*, vol. 24, 2006, pp. 1–24.

[Kod99] Kodratoff Y. Quelques contraintes symboliques sur le numérique en ecd et ect. *In : SFDS*, pp. 183–188, 1999. Grenoble,France.

[Lal02] Lallich S. Mesure et validation en extraction des connaissances à partir des données, 2002. Thèse d'habilitation à Diriger les Recherches, Université Lyon 2.

[Ler81] Lerman I.C. *Classification et analyse ordinale des donnés*. Dunod, 1981.

[Ler84] Lerman I. C. Justification et validité statistique d'une échelle [0,1] de fréquence mathématique pour une structure de proximité sur un ensemble de variables observées, 1984. Rapport de recherche INRIA, Centre de Rennes IRISA.

[LFZ99] Lavrac N., Flach P. et Zupan B. Rule evaluation measures : A unifying view. *In : Ninth international workshop on Inductive Logic Programming*, éd. par Mineau G. et Ganter B., pp. 174–185, 1999.

[LGR81] Lerman I.C., Gras R. et Rostam H. Elaboration et évaluation d'un indice d'implication pour des données binaires. *Math Sc. Hum*, vol. 74, 1981, pp. 5–35.

[LHCM00] Lui B., Hsu W., Chen S. et Ma Y. Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems*, vol. 15(5), 2000, pp. 47–55.

[LK98] Lin Dao-I et Kedem Zvi M. Pincer search: A new algorithm for discovering the maximum frequent set. *Lecture Notes in Computer Science*, vol. 1377, 1998, pp. 105–121.

[LMV+04] Lenca P., Meyer P., Vaillant B., Picouet P. et Lallich S. Evaluation et analyse multi-critère des mesures de qualité des règles d'association. *RNTI-E-1*, 2004, pp. 219–246.

[LMVP03] Lenca P., Meyer P., Vaillant B. et Picouet P. Aide multicritères à la décision pour évaluer les indices de qualité de connaissances. *In : Proc. of the EGC Conference*, pp. 271–282, 2003. Lyon,France.

[Loe47] Loevinger J. A symetric aproach to the construction and evaluation of tests of ability. *Psychological Monographs*, vol. 61, 1947, pp. 1–49.

[LT04] Lallich S. et Teytaud O. Evaluation et validation de mesures d'intérêt des règles d'association. *RNTI-E-1*, vol. spécial, 2004, pp. 193–217.

[Luo06] Luong V.P. Reasoning on association rules, 2006. Rapport technique. Labo Informatique de Marseilles.

[LYKC02] Loo K. K., Yip C. Lap, Kao B. et Cheung D. A lattice-based approach for I/O efficient association rule mining. *Information Systems*, vol. 27, 2002, pp. 41–74.

[MM93] Major J. A. et Mangano J. Selecting among rules induced from a heuricane database. *In : KDD Workshop papers, Menlo Park California*, pp. 28–41, 1993.

[Mon03] Monjardet B. The presence of lattice theory in discrete problems of mathematical social sciences. why. *Mathematical Social Sciences*, vol. 46, 2003, pp. 103–144.

[Mor62] Morgado J. A characterization of the closure operator by mean of one axiom. *Portugal Math*, vol. 21, 1962, pp. 155–156.

[MT97] Mannila H. et Toivonen H. Levelwise search and borders of theories in knowledge discovery. *Data Mining Knowledge Discovery*, vol. 1. 3, November 1997, pp. 241–258.

[Ö44] Öre O. Galois connections. *Transaction of the American Mathematical Society*, vol. 55, 1944, pp. 494–513.

[Pas00] Pasquier N. *Data Mining : Algorithmes d'extraction et de réduction des règles d'associationdans les bases de données.* Clermont-Ferrand, FRANCE, Thèse de PhD, Clermond-Ferrand II, 2000.

[PBTL99] Pasquier N., Bastide Y., Taouil R. et Lakhal L. Efficient mining of association rules using closed itemset lattices. *Information Systems*, vol. 24, 1999, pp. 25–46.

[PHM00] Pei J., Han J. et Mao R. CLOSET: An efficient algorithm for mining frequent closed itemsets. *In : ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pp. 21–30, 2000. Dallas,U.S.A.

[Plo73] Plott C.R. Path independance, rationality and scocial choice. *Econometrica*, vol. 41, 1973, pp. 1075–1091.

[Pol98] Polaillon G. *Organisation et interprétation par les treillis de Galois de données de type multivalué, intervalle ou*

histogramme. France, Thèse de PhD, Université Parix IX-Dauphine, 1998.

[PS91a] Piatetsky-Shapiro G. Discovery, analysis, and representation of strong rules. *Knowlegde Discovery in Databases*, vol. AAAI Press/The MIT Press, 1991, pp. 229–248.

[PS91b] Piatetsky-Shapiro G. Knowledge discovery in real data bases. *AI Magazine*, vol. 11, 1991, pp. 68–70.

[SA96] Srikant R. et Agrawal R. Mining sequential patterns: Generalizations and performance improvements. *In : Proc. of 5th Biennial International Conference on Extending Database Technology (EDBT'96)*, pp. 3–17, 1996. Avignon,France.

[SS88] Sebag M. et Shoenauer M. Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases. *In : Proc. of the European Knowledge Acquisition Workshop Conference*, pp. 28–1–28–20, 1988. Bonn,Germany.

[STB+01] Stumme G., Taouil R., Bastide Y., Pasquier N. et Lakhal L. Intelligent structuring and reducing of association rules with formal concept analysis. *In : Advances in Artificial Intelligence*, éd. par Baader F., Brewka G. et Eiter T. pp. 335–350, 2001. Springer-Verlag.

[STB+02] Stumme G., Taouil R., Bastide Y., Pasquier N. et Lakhal L. Computing iceberg concept lattices with TITANIC. *Data and Knowledge Engineering*, vol. 42, 2002, pp. 189–222.

[Tot92] Totohasina A. *Méthode implicative en Analyse de données et Application l'analyse de conceptions d'tudiants sur la notion de probabilité conditionnelle*. France, Thèse de PhD, Université de Rennes I, 1992.

[Tot94] Totohasina A. *Notes sur l'implication statistique: dépendance positive orientée, valeurs critiques*. Rapport technique, Université du Québec à Montréal, SCAD, Dept de Maths-Info, 1994.

[Tot03] Totohasina A. Normalisation de mesures probabilistes de la qualité des règles. *In : Proc. SFDS'03, XXXV ième Journées de Statistiques*, pp. 985–988, 2003. Lyon 2, France.

[Tot06] Totohasina A. Extraction des règles d'association à sémantique d'implication à partir des données quantitatives. *In : Proc. of XXXVIIIemes journées de Société Franaise de Statistique, SFDS'06, 29 mai-02 juin*, p. 6 pages, 2006. Clamart, France.

[TR05] Totohasina A. et Ralambondrainy H. Ion : a pertinent new measure for mining information from many types of data. *In : IEEE SITIS'05.*, pp. 202–207, 2005. Yaoundé, Cameroun.

[TRD03] Totohasina A., Ralambondrainy H. et Diatta J. Un algorithme efficace d'extraction des règles d'association implicative, Decembre 2003. Rapport technique de Recherche.

[TRD04] Totohasina A., Ralambondrainy H. et Diatta J. Notes sur les mesures probabilistes de la qualité des règles d'association : un algorithme efficace d'extraction des règles d'association implicative. *In : Proc. of CARI'04*, pp. 511–518, 2004. Hammamet, Tunisie.

[TRD05] Totohasina A., Ralambondrainy H. et Diatta J. Une vision unificatrice des mesures probabilistes de la qualité des règles d'association booléennes et un algorithme efficace d'extraction des règles d'association implicative. *In : Proc. of TAIMA'05*, pp. 375–380, 2005. Hammamet, Tunisie.

[Vai06] Vaillant B. *Mesurer la qualié des règles d'association : études formelles et expérimentales*. France, Thèse de PhD, École Nationale Supérieure des Télécommunications de Bretagne et Université de Bretagne sud, 2006.

[Val99] Valtchev P. *Construction automatique de taxonomies pour l'aide la reprsentation des connaissances par objets*. France, Thèse de PhD, Université de Grenoble 1, 1999.

[Vou02] Voutsadakis G. Polyadic concept analysis. *Order*, vol. 19, 2002, pp. 295–304.

[Web01] Webb Geoffroy I. *Discovering associations with numeric variables*. Rapport technique Geelong-Vic-3217, Australia, School of computing and Mathematics, 2001.

[Wil82] Wille R. Restructuring lattice theory: an approach based on hierarchies of concepts. *In : Ordered sets*, éd. par Rival I., pp. 445–470. Dordrecht-Boston, Ridel, 1982.

[WZZ04] Wu X., Zhang C. et Zhang S. Efficient mining of both positive and negative association rules. *ACM Transactions on information Systems*, vol. 3, 2004, pp. 381–405.

[ZH99] Zaki M. J. et Hsiao C.-J. CHARM: An efficient algorithm for closed itemset mining, 1999. Technical Report 99-10, Computer Science, Rensselaer Polytechnic Institute, 1999.

[Zha00] Zhang T. Association rules. *In : PAKDD 2000*. pp. 245–256, 2000. Springer-Verlag.

[ZO98] Zaki M. J. et Ogihara M. Theoretical Foundations of Association Rules. *In : 3rd SIGMOD'98 Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD)*, pp. 1–8, 1998.

[ZPOL97] Zaki M. J., Parthasarathy S., Ogihara M et Li W. New algorithms for discovery association rules. *In : Knowledge Discovery and Data Mining*, pp. 283–296, 1997.

Liste des figures

Liste des tableaux

[10pt,a4paper]article [francais]babel [latin1]inputenc

Chapitre 6

Annexe : Dossier personnel

Curriculum vitae

TOTOHASINA André

- **Adresse personnelle :**
  Lot 400 ALE, Cité des professeurs
  Lazaret Sud. Antsiranana - 201
  MADAGASIKARA
- **Adresse professionnelle :**
  ENSET. Université d'Antsiranana
  B.P. O
  Antsiranana - 201
  MADAGASIKARA

  Tél. : 0320790923(00261320790923)
  e-mail : totohasina@yahoo.fr

État civil

Né le 04-06-1954 à Mandritsara, Madagasikara.
Marié, 4 enfants
Nationalité MALAGASY
Chevalier de l'Ordre National.

Profession

Enseignant chercheur, Maître de conférences, 1ère classe, 3ème éch.
Spécialités : Mathématiques et Didactique des mathématiques.

Études

1973–1975	Baccalauréat C (mathématiques et sciences physiques). *École normale des instituteurs de Mahamasina Antanànarivo, Madagasikara*
1976–1977	Certificat d'aptitude pédagogique pour le CEG (CAP-CEG) en Mathématiques et Sciences physiques, Major de la promotion, Nov. 1978. *INSRFP d'Ampefiloha Antanànarivo, Madagasikara.*
1980–1985	CAPEN en Mathématiques,Major de la promotion, Nov. 85. *École Normale Supérieure(ENN3) de Fianarantsoa, Madagasikara.*
1988–1989	DEA de mathématiques option didactique, mention Assez-Bien, Septembre 89. *IRMA, Université Louis Pasteur I, France.*
1991–1992	Cours de DEA de Probabilités et de Prparation à l'agrégation de mathématiques, option Probabiliés et statistiques. *IRMAR de l'Université de Rennes I, France.*
1989–1992	Doctorat nouveau régime, spécialité Mathématiques et Applications, mention très Honorable, 27/11/92. *IRMAR de l'Université de Rennes I, France.*

Stages de perfectionnement

Avril 2007	Épidémiologie ; Modèles statistiques en épidémiologie ; Apprentisage statistyique et data mining ; Markoc Chain Monte Carlo Method(MCMC) ; Modèle linéaire mixte et algorithme EM ; Analyse de Survie : techniques pour des données censurées, méthodes d'estimation non paramétrique ; Le modèle de Coxe ; Manipulation du logiciel R. *École CIMPA(Centre international de mathématiques pures et appliquées) de statistiques mathématiques et statistique de la Santé, Université de Yaoundé I, Cameroun.*
Mars 2006	Le système LINUX pour administration de réseau informatique.*AUF/IST-Diégo.*
Septembre 2005	Formation intensive : Tutorer une formation à distance sur la plateforme acolad.*AUF/Campus numérique francophone de l'océan indien, Tsimbazaza, Antanànarivo.*
Septembre 2004	Multimédia.*MADSUP/Unité de production multimédia, Université d'Antanànarivo.*
Fév. 03	Stage de recherche en didactique des mathématiques. *IREM, Université Louis Pasteur de Strasbourg I, France.*
Octobre 2002	Utilisation des logiciels d'analyse de données : SPAD et SPSS.*MADSUP/Faculté des Sciences, Université d'Antanànarivo.*

Expériences professionnelles en recherche

Oct. –déc. 07	Finition des travaux pour HDR en Mathématiques et Informatique. *Équipe ECD, laboratoire IREMIA, Faculté de Sciences et technologies, Université de La Réunion, France.*
Avr. - Mai 06	Maître de conférences invié : Recherche et Enseignement de Processus stochastiques en Licence d'Informatique 3 et IUP2, Faculté de Sciences et technologies. *Université de La Réunion, France.*
Mai. - Juillet 05	Maître de conférences invié : Recherche et Enseignement d'Analyse 2 en Maths et informatique Appliquée L2, Faculté de Sciences et technologies. *Université de La Réunion, France.*
Oct. 03–Janv. 04	Stage de recherche en fouille des données en vue d'HDR. *Équipe ECD, laboratoire IREMIA, Faculté de Sciences et technologies, Université de La Réunion, France.*
Oct. 02–Janv. 03	Stage de recherche en fouille des données en vue d'HDR. *Équipe ECD, laboratoire IREMIA, Faculté de Sciences et technologies, Université de La Réunion, France.*
Sept. 93–Juill. 94	Stage postdoctoral au CIRADE(Centre interdisciplinaire de recherche appliquée au dévéloppement de l'enfant) comme personne ressource en traitements et analyse des données et au SCAD (Service de Conseil en analyse des Données), équipe de statistique du département de mathématiques, *Université du Québec à Montréal, Canada.*

Expériences professionnelles en service collectif et /ou Poste occupé en administration universitaire

2002-2005	Superviseur de stage pédagogique et Chef de département de mathématiques. *ENSET, Université d'Antsiranana.*
1995-2002	Directeur de l'École normale supérieure (ENSET). *Université d'Antsiranana.*
1996–1998	Responsable locale d'un DEA de maths, option Analyse et statistiques appliquées. *Université d'Antsiranana.*
1994–1995	Maître de conférences : prof. de mathématiques (Analyse I, II et probabilités et statistiques) à la faculté de sciences. *Université d'Antsiranana.*
1985–1988	Professeur certifié CAPEN : professeur de mathématiques.*Lycée d'Antsohihy.*
1978–1980	Chargé d'Enseignement CAP/CEG : professeur de mathématiques et sc. physiques, *CEG de Mahamanina, Fianarantsoa.*

Enseignements universitaires annuels effectués en ET et ED/ EP

1996– 2007 À l'ENSET :
- Première année : Analyse I, Algèbre I, Statistique I.
- PETM / GMI 2 : Analyse 2, Probabilités et statistique, Algèbre, géométrie affine et projective.
- PETM/GMI 3 : Algèbre multilinéaire et tenseurs, compléments de probabilités et statistiques, programmation linéaire.
- PETM/GMI 4 : Algèbre (Arithmétique, Anneaux et extensions de corps), Introduction aux processus stochastiques, Analyse fonctionnelle, Didactique et pédagogie des mathématiques, EDPs et transformations intégrales.

Université d'Antsiranana.

2007 Mathématiques IV (65h). PC4, option physique. *Faculté des sciences. Université d'Antsiranana.*

2005–2007 Modélisations mathématiques (déterministes et stochastiques). DEA de système énergétique et mécaniques des fluides. *Faculté des sciences.*

2005–2007 Compléments de mathématiques (transformations et équations intégrales, optimisations), 50h ET. DEA d'électromécanique. *École supérieure polytechnique. Université d'Antsiranana.*

2005–2008 Caculs intégrales et équations différentielles GTR-1(35h), Maths pour signal discret (35h) GTR-2, Maths pour informatique et réseaux (30h) GTR-2. *IST-Antsiranana.*

1996 Statistiques mathématiques(37,5h) et Analyse de données(37,5h). DEA de mathématiques. *Université d'Antsiranana.*

Thèmes de recherche et publications

A. Thèmes de recherche : Fouille de données, et Didactique des mathématiques et Formation des enseignants.

B. Publications :

B-1. Sur la fouille de données et applications. B-1.1. Dans des revues spécialisées de notoriété internationale

- 1. R. Gras & A. Totohasina (1995), Conceptions d'élèves sur la notion de probabilité conditionnelle révélées par une méthode d'analyse des données : implication-similarité-corrélation, in Revue Educational Studies in Mathematics 28, Kluwer Academic Publishers, Printed in the Netherlands, 1995, 337-363.
- 2. R. Gras & A. Totohasina (1995), Chronologie et causalité, concep-

tions sources d'obstacles épistémologiques à la notion de probabilité conditionnelle, in revue Recherche en Didactique des Mathématiques, Vol.15, n¡1, La Pensée Sauvage (édts), Grenoble, France, 1995, 49-95.

– 3. Feno D, Diatta J., Totohasina A.(2007), *Une base pour les règles d'association valides au sens de la mesure de qualité* M_{GK}, in Revue de la Nouvelle Technologie de l'Information, RNTI, issue spéciale de SFC'2006, version longue, 11 pages (à paraître).
– 4. Feno D., Diatta J., Totohasina A.(2007), *Galois lattices and Bases for* M_{GK}*-valid association rules*, Revisited version, in Lecture Note in Artificial Intelligence, Belohlavek & al editors, special issue of CLA 2006 (à paraître).
– 5. Diatta J., Feno D., Totohasina A.(2007), *Mining Bases for* M_{GK}*-valid association rules*, Revisited version, in Global Journal for Pure and Applied Mathematics, GJPAM, Research India Publications, (9 pages)(Accepted to appear).

B-1. 2. Ouvrages collectifs :

– 6. Réponse à un appel international à chapitres de livre :
Diatta J., Ralambondrainy H., André Totohasina (Janvier 2007), *Towards a unifying Implicative Normalized probabilistic quality measure for association rules*, in Quality Measure in Data Mining, Series Studies in computational intelligence, Guillet Fabrice & Hamilton Howard editorsVol 43, 10^{th} Chapter , 237-238.

– 7 Initiative du groupe de chercheurs concernés :
R. Gras, S. Ag Almouloud, A. Larher, H. Ratsimba, A. Totohasina (1996), *L'Implication statistique . Une nouvelle méthode d'Analyse exploratoire*, La Pensée sauvage (éditions), Grenoble, France, 1996.

B-1. 3. Sur des actes de colloques internationaux :

– 8. Totohasina A. (2003) *Normalisation des mesures probabilistes de la qualité des règles d'association*, in Proceedings Société Statistique de France XXXVè Journées SFDS, Université Lumière Lyon 2, France, 2003, pp. 985-988.

– 9. Totohasina A., Ralambondrainy H., Diatta J. (2004), *Notes sur les mesures probabilistes de la qualité des règles d'association : un algorithme efficace d'extraction des règles d'association implicative*, Proceedings Colloque Africain sur la Recherche en Informatique, CARI, Hammamet, Tunisie, 2004, 511-512.
– 10. Totohasina A., Ralambondrainy H. , Diatta J.(2005), *Une vision unificatrice des mesures probabilistes de la qualité des règles d'association booléenes et un algorithme efficace d'extraction des règles d'association implicative*, proceedings of Atelier francophone

de Traitement et Analyse de l'Information : Méthodes et Applications, TAIMA 2005, Hammamet, Tunisie, 26 septembre-1er octobre 2005, 375-380.

- 11. Totohasina A., Ralambondrainy H. (2005), *ION : a pertinent new measure for mining information from many types of data*, proceedings of The 2005 International Conference on Signal-Image Technology & Internet- Based Systems (SITIS'05), November 27th - December 2nd 2005, The Hilton Hotel, Yaoundé, Cameroon, 202-207.
- 12. Feno D., Diatta J., Totohasiina A.(2006), *Normalisée d'une mesure probabiliste de la qualité des règles d'association : étude de cas*, Proceedings of EGC 06, Qualité des données et des Connaissances (QDKQ), Lille, France, 2006, 25-30.
- 13. Feno D., Diatta J., Totohasina A.(2006), *Une base pour les règles d'association valides au sens de la mesure de qualité* $\mathrm{M_{GK}}$, Proceedings of XIIIme Rencontres SFC 06, Socié Française de Classification, Metz, France, 2006.
- 14. Feno D., Diatta J., Totohasina A.(2006), *Génération des bases pour les règles d'association* $\mathrm{M_{GK}}$*-valides*, Proceedings of XIVmes Rencontres SFC 07, Socié Française de Classification, TELECOM Paris, 5-7 Septembre 2007, 101-105.
- 15. Feno D., Diatta J., Totohasina A.(2006), *Galois lattices and Bases for* $\mathrm{M_{GK}}$*-valid association rules*, Long papers, in Procceedings of4^{th} International Conference on Concept Lattices and Their Applications, CLA 06, Hamamet, Tunisie, october 30-november 1st, 2006, 127-138.
- 16. Totohasina A. (2005), *Une nouvelle méthode d'extraction des règles d'association quantitative*, Proceedings of Atelier francophone de Traitement et Analyse de l'Information : Méthodes et Applications, TAIMA 2005, Session invitée Analyse Symbolique de l'Information ASI, Hammamet, Tunisie, 26 septembre-1er octobre 2005, 54-59.
- 17. Totohasina A.(2006), *Extraction des règles d'association à sémantique d'implication à partir des données quantitatives*, Proceedings of XXXVIIIemes journées de Société Française de Statistique, SFDS 2006, 29 mai-2 Juin 2006, Clamart, France, 2006.
- 18. Diatta J., Ralambondrainy H., Totohasina A.(2007), *Règles d'association dans un contexte de descriptions ordonnées*, Proceedings of XIVmes Rencontres SFC 07, sociétété Française de Classification, TELECOM Paris, 5-7 Septembre 2007, 94-96.

B-1. 3. Sur des actes de colloques nationaux : .

- 19. Totohasina A. (avril 2005), *L'implication statistique orientée normalisée (ION) : vers un outil de la fouille d'un grand volume des données de divers types*, Actes de Forum de Recherches du MENRS, Toamasina, 22 pages, 2005.

– 20. Totohasina A. (2000), *Étude d'une situation didactique en statistique double par l'implication statistique*, Journées de la recherche du MINSUP, Fianarantsoa, 20 pages, Juin 2000.
– 21. Totohasina A. et Feno D. R. (1999), *Suggestions sur la pédagogie de résolution de problèmes. Taxonomie à partir de l'implication statistique de Gras*, Actes du colloque international sur la didactique des disciplines, ENS d'Antanànarivo, 12 pages, 1999.

B-1. 3. Articles soumis à un colloque international (Février 2008) :

– (S1). Totohasina A. (2008), *De la qualité des règles d'association. Une normalisation unificatrice des mesures, rôle de* M_{GK}, 8 pages.(soumis).

– (S2). Totohasina A., Feno D. R.(2008), *De la qualité des règles d'association : Étude comparative des mesures* $\mathrm{M_{GK}}$ *et* $Confiance$. (8 pages) (soumis).

B-1. 3. Preprint / Rapports techniques de recherche non publiés :

– (R1). Totohasina A. (1994), *Notes sur les valeurs critiques de l'indice d'implication de Gras*, Rapport technique de recherche, Équipe statistique, Département de mathématiques, Université du Québec à Montréal, Canada (non publié), 28 pages.
– (R2). Totohasina A., Ralambondrainy H., Diatta J. (2003), *Un algorithme efficace d'extraction des règles d'association implicative*, Rapport technique de recherche, équipe ECD, Laboratoire IREMIA, Dépt. de maths-Info., Faculté des sciences et Technologies, Université de La Réunion, France, 22 pages.
– (R3). Totohasina A. (2004), *Une mesure de qualité des règles d'association quantitative*, équipe ECD, Laboratoire IREMIA, Dépt. de maths-Info., Faculté des sciences et Technologies, Université de La Réunion, France, 19 pages.

Sur la didactique des mathématiques et la formation des enseignants

– (Did1). Totohasina A. (Avril 2005), *Un plaidoyer aux pertinence et faisabilité de l'introduction précoce des coniques*, Actes de Forum de Recherches du MENRS, Toamasina, 22 pages, 2005.
– (Did2). Totohasina A. (Avril 2005), *Au niveau scolaire (collège et lycée) les nombres qualifiés de "racines évidentes" d'un polynôme à coefficients rationnels sont-ils vraiment évidents ?*, Actes de Forum de Recherches du MENRS, Toamasina, 18 pages, 2005.
– (Did3). Totohasina A. (2004), *À propos de l'apprentissage des formes et constructions géométriques dans un plan : Un plaidoyer aux per-*

tinence et faisabilité de l'introduction précoce des coniques, in revue DIDAKTIKA , Vol. 2, Centre Interuniversitaire de Recherche en Didactique, CIRD, ENS d'Antanànarivo, 2004.

- (Did4). Totohasina A., Rakotondrasoa H. et Rabe Tsirobaka(2004) *Vers la construction d'un modèle de dispositif d'encadrements de stage pédagogique sur des disciplines scientifiques et techniques*, in revue DIDAKTIKA , Vol. 2, Centre Interuniversitaire de Recherche en Didactique, CIRD, ENS d'Antanànarivo, pp. 63-80, 2004.
- (Did5). Totohasina A. (1994), *Une méthode d'introduction des probabilités conditionnelles : avantages et inconvénients de l'arborescence*,In revue Repères -IREM, Topiques - Editions, France, Vol. 15, pp. 93-117.
- (Did6). Totohasina A. (1994), *L'implication statistique en classification*, SCAD, Dépt de maths et informatique, Université du Québec à Montréal, Canada, 6 pages. (Conf. sans comité de lecture).
- (Did7). Totohasina A. (1994), *Conceptions d'étudiants et d'étudiantes universitaires sur la notion de probabilité conditionnelle*, Séminaire du CIRADE, Université du Québec à Montréal, Canada, 6 pages. (Conf. sans comité de lecture).
- (Did8) Totohasina A. (1993), *Les analyses factorielles (à la française) : une présentyation géérique des méthodes classiques ACP, AFC, AFM, AFD*, Séminaire du SCAD, Dépt de maths et info., Université du Québec à Montréal, Canada, 6 pages. (Conf. sans comité de lecture, 9 octobre 1993).
- (Did9). Totohasina A. (1993), *Quelques misconceptions qui risquent de devenir obstacles épistémologiques sur la notion de probabilité conditionnelle*, in Fascicule de didactique des mathématiques, Institut de Recherche Mathématique de Rennes I, France, 43 pages. (Conf. sans commité de lecture).
- (Did10). Totohasina A. (1993), *Quelques conceptions d'étudiants sur la notion de probabilité conditionnelle*, Séminaire de didactique de mathématiques, Dépt de maths et info., Université du Québec à Montréal, Canada, 6 pages. (Conf. sans comité de lecture, 6 octobre 1993).

Encadrements de thèse, DEA de mathématiques (soutenus)

Déc. 2007	Doctorat de l'université de La Réunion : *Mesures de qualité pour les règles d'association : normalisation et caractérisation de bases.*, Feno Daniel Rajaonasy.
Avril 1999	DEA : *Position de l'implication statistique de Gras en tant que concept de dépendance statistique par rapport à la classification de Lehmann et application en didactique des mathématiques*, Feno Daniel Rajaonasy. (Université d'Antsiranana, en partenariat avec Faculté des sciences d'Antanànarivo)

Encadrements de DEA en cours : Système énergétique et mécaniques de fluides.

Février 2008	Ambeondahy, *Résolution numérique d'une équation de chaleur à deux dimensions par la méthode des éléments finis et comparaison avec solutions analytiques*, Faculté des Sciences, université d'Antsiranana.
Février 2008	Pascal Petera,*Modélisation statistique d'un problème d'écoulement d'un fluide sur une plaque chauffée*, Faculté des Sciences, université d'Antsiranana.

Encadrements de Mémoires de fin d'études et de projets en vue de CAPEN.

07 fév. 2008 *Fonction de Répartition*, Mizdali, GMI, ENSET, Université d'Antsiranana.

07 fév. 2008 *Estimation statistique par le Jacknife*, Ratsimbarison Frédien Jacques, GMI, ENSET, Université d'Antsiranana.

2005 *Programmation linéaire perturbé et exploitation pédagogique d'Excel via Solver.*, Ramanantsoa Harriman, GMI, ENSET, Université d'Antsiranana.

2005 *Apprentissage des méthodes de résolution d'équations algébriques du collège au lycée.*, Mohammed Attouman, GMI, ENSET, Université d'Antsiranana.

2005 *Pertinence et faisabilité de l'Apprentissage des quadriques en seconde*, Ratiambelo, GMI, ENSET, Université d'Antsiranana.

2005 *Apprentissage des moyennes conditionnelles en secondaire et de l'espérance conditionnelle en supérieur.*, Crisse Jean, GMI, ENSET, Université d'Antsiranana.

2005 *Introduction intuitive des intégrales multiples*, Nadhoiri Ali, GMI, ENSET, Université d'Antsiranana.

2004 *Faisabilité de l'introduction des lois de probabilités continues en terminale.*, Masondrazana, GMI, ENSET, Université d'Antsiranana.

2004 *Apprentissage des calculs mentaux et astuces rapides en mathématiques.*, Be Jérôme, GMI, ENSET, Université d'Antsiranana.

2004 *Apprentissage de la résolution de problèmes au niveau du second cycle secondaire.*, Ronto Sidoni, GMI, ENSET, Université d'Antsiranana.

2004 *Élaboration d'un logiciel de résolution d'un programme linéaire par la méthode de simplexe(Projet).*, Ramanantsoa Harriman, GMI, ENSET, Université d'Antsiranana.

2004 *Élaboration d'un logiciel de quadriques sur Matlab (Projet).*, Ronto Sidoni, GMI, ENSET, Université d'Antsiranana.

2004 *Élaboration d'un logiciel de quadriques sur Matlab (Projet).*, Ratiambelo, GMI, ENSET, Université d'Antsiranana.

2004 *Élaboration d'un logiciel d'implication statistique selon Totohasina & al. sur Matlab(P)* , Armand, GMI, ENSET, Université d'Antsiranana.

2003 *Élaboration d'un logiciel de tests statistiques non paramétriques sur Matlab (P).*, Armand, GMI, ENSET, Université d'Antsiranana.

2003 *Algorithme de recherche des racines rationnelles d'un polynôme de Q[X] (P)*, Masondrazana, GMI, ENSET, Université d'Antsiranana.

2003 *À propos des nombres complexes : interprétations géométriques des opérations algébriques, transformations conformes et applications en physiques(P)*, Ronto Sidoni, GMI, ENSET, Université d'Antsiranana.

2003 *Sur l'analyse des transformations intégrales : exemples d'applications en physique et extension en deux variables (P)*, Auberto, GMI, ENSET, Université d'Antsiranana.

Encadrements de Mémoires (suite)

2003 *Sur l'analyse des transformations intégrales : exemples d'applications en physique et extension en deux variables.(projet)*, Auberto, GMI, ENSET, Université d'Antsiranana.

2003 *Arithmétique de $\mathbb{Z}$: à propos du théorème chinois des restes et algorithme de résolution(Algèbre)*, Baovavelo Natacha, GMI, ENSET, Université d'Antsiranana.

2003 *À propos des méthodes statistiques d'analyse des données*, Rakotovao Félicia Romie, GMI, ENSET, Université d'Antsiranana.

2000 *Activités d'introduction de quelques concepts en mathématiques,* Soaziliny, GMI, ENSET, Université d'Antsiranana.

2000 *Élaboration d'un logiciel de statistique au niveau secondaire*, Ndriamaharo Hery, GMI, ENSET, Université d'Antsiranana.

2000 *Élaboration d'exercices corrigés et d'un précis de cours en Probabilités et Statistiques (niveau secondaire)*, Robiseheno, GMI, ENSET, Université d'Antsiranana.

1999 *Méthodes numériques en E.D.P.*, Marotsara Jean Hugo, GMI, ENSET, Université d'Antsiranana.

1999 *Femmes et mathématiques à Diégo-Suarez*, Zaramanana Céleste, GMI, ENSET, Université d'Antsiranana.

1999 *Enseignement de l'Arithmétique au Lycée*, Salim Abdou Salim, GMI, ENSET, Université d'Antsiranana.

1999 *Enseignement de la statistique double au Lycée*, Razanatsoa Célestin, GMI, ENSET, Université d'Antsiranana.

1998 *Réalisation d'un logiciel d'analyse numérique*, Andriamampionona Manitriniaina, GMI, ENSET, Université d'Antsiranana.

1998 *Construction d'un polygone régulier à n côtés constructibles ou non. (Algèbre)*, Ravalison Justin, GMI, ENSET, Université d'Antsiranana.

1996 *Les transformations planes et les pavages réguliers.*, Randrianome Harinelina, GMI, ENSET, Université d'Antsiranana.

1996 *La logique et ses applications*, Ranaivosolo patrick William, GMI, ENSET, Université d'Antsiranana.

1996 *Équilibre théorie - pratique en mathématiques*, Ramaroson Rodrigues, GMI, ENSET, Université d'Antsiranana.

1995 *Équations différtentielles du 1er et du 2ième ordre. Quelques applications*, Kaissany, GMI, ENSET, Université d'Antsiranana.

1995 *Réflexion didactique sur la notion de probabilité en articulation avec la statistique descriptive*, Youssouf M'madi, GMI, ENSET, Université d'Antsiranana.

1995 *Réflexion didactique sur le concept de probabilité conditionnelle*, Feno Daniel R., GMI, ENSET, Université d'Antsiranana.

Expertises

2007–2008	Comité du programme du colloque africain sur la recherche en informatique et en mathématiques appliquées du 27-30 octobre 2008, Rabat, Maroc (CARI'08) : comité de lecture dans le thème 5 du Système d'information.
2007–2008	Projet de recherche **INDIMAR** : Partenaire numéro 6 et Responsable de l'extraction des connaissances à partir de données satellitales, Analyse implicative, classification hiérarchique implicative et cohésitive, fouille de données, en vue d'**identifier les concepts discriminants**. IRD et université de La Réunion.
2000–2006	Co-fondateur du Centre interuniversitaire de recherche en didactique (CIRD), ENS d'Antanànarivo.

Notes sur le Projet de recherche INDIMAR : Le projet de recherche INDIMAR a été initié par l'équipe ECD de l'IREMIA de l'Université de La Réunion et la station SEAS-Réunion de l'Institut de Recherche pour le Développement (IRD). Il ambitionne de créer un lien pratique entre la recherche mathématique - informatique et la recherche méthodologique de l'axe O.T. de l'US 140 de l'IRD.
L'objectif de ce projet est d'élaborer des outils de détection et de suivi automatiques d'indicateurs oéaniques, à partir des cartes de température de surface de la mer (carte SST) résultant de la chaîne de traitement des images satellites NOAA, réceptionnées par la station SEAS-Réunion.
On s'attend que les résultats de ce projet contribueront alors à la surveillance de l'environnement marin de la zone Océan indien, notamment pour le développement de la pêche durable.
Les résultats attendus se déclinent sous forme de trois indicateurs majeurs :

- (A) au co-développement des Pays de la zone Océan Indien : les étudiants de la zone, notamment des malagasy du DEA d'informatique, y seront encadrés en formation doctorale.
- (B) à la coopération régionale et partenariats sud-sud via avec les universités d'Antsiranana et de Tunis.
- (C) à la coopération intra-France (Université d'Orléan), régionale (Université d'Antsiranana) et internationale (Birkbeck college, University of London, faculté des sciences de Tunis).

Projet en instance STAFAV au sein de SARIMA-Madagasikara

2007–2008 Création à l'Université d'Antsiranana d'un Master de Statistique Appliquées à la science du vivant dont l'épidémiologie pour adhérer et étendre le champ du réseau de Masters à double diplômes de STAFAV initié par un groupe d'éminents probabilistes, statisticiens, épidémiologistes français (Projet FSP du MAE de la France).

Sociétés savantes

2007– RASMA : Réseau africain de statistiques et mathématiques appliquées.

2005– SARIMA-CIMPA : Soutien africain de recherche en informatique et mathématiques appliquées - Centre international de mathématiques pures et appliquées.

1990–1995 SFC : Société française de classification.

1990–1995 ARDM : Association de recherche en didactique de mathématiques.