Mesurer la qualité des règles d'association

Benoît Vaillant

▶ **To cite this version:**

Benoît Vaillant. Mesurer la qualité des règles d'association : études formelles et expérimentales. Mathématiques [math]. Ecole nationale supérieure des télécommunications de Bretagne, 2006. Français. NNT : 2006TELB0026 . tel-03247307

HAL Id: tel-03247307

https://imt-atlantique.hal.science/tel-03247307

Submitted on 3 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N° d'ordre : 2006telb0026

Thèse

Présentée à

l'École Nationale Supérieure des Télécommunications de Bretagne

en habilitation conjointe avec l'Université de Bretagne Sud

pour obtenir le grade de

Docteur de l'ENST Bretagne
Mention « **Mathématiques et Informatique** »

par

***Benoît* Vaillant**

« *Mesurer la qualité des règles d'association : études formelles et expérimentales* »

Soutenue le 5 Décembre 2006 devant la Commission d'Examen :

Composition du Jury

- *Président* : Howard Hamilton, Professeur, Université de Regina
- *Rapporteurs* : Pascale Kuntz, Professeur, Ecole polytechnique Nantes
  Djamel Abdelkader Zighed, Professeur, Université Lumière Lyon 2
- *Examinateurs* : Pierre-François Marteau, Professeur, Université de Bretagne Sud
  Alain Hillion, Professeur, ENST Bretagne
  Philippe Lenca, Maître de Conférences, ENST Bretagne

N° d'ordre : 2006telb0026

Thèse

Présentée à

l'École Nationale Supérieure des Télécommunications de Bretagne

en habilitation conjointe avec l'Université de Bretagne Sud

pour obtenir le grade de

Docteur de l'ENST Bretagne
Mention « **Mathématiques et Informatique** »

par

***Benoît* Vaillant**

« *Mesurer la qualité des règles d'association : études formelles et expérimentales* »

Soutenue le 5 Décembre 2006 devant la Commission d'Examen :

Composition du Jury

- *Président* : Howard Hamilton, Professeur, Université de Regina
- *Rapporteurs* : Pascale Kuntz, Professeur, Ecole polytechnique Nantes
  Djamel Abdelkader Zighed, Professeur, Université Lumière Lyon 2
- *Examinateurs* : Pierre-François Marteau, Professeur, Université de Bretagne Sud
  Alain Hillion, Professeur, ENST Bretagne
  Philippe Lenca, Maître de Conférences, ENST Bretagne

Remerciements

Je tiens à exprimer toute ma reconnaissance à Alain Hillion pour avoir accepté de diriger mes recherches, ainsi qu'à Philippe Lenca pour son encadrement, ses nombreux conseils et son soutien constant tout au long de ma thèse.

Je remercie Pascale Kuntz et Djamel Zighed d'avoir accepté d'être rapporteurs de mes travaux. Je tiens à leur exprimer ma profonde gratitude pour leurs commentaires et jugements très pertinents sur mon mémoire, tant sur le fond que sur la forme.

Je tiens également à assurer Howard Hamilton et Pierre-François Marteau de ma reconnaissance pour avoir accepté de faire partie de mon jury de thèse. Je suis fortement touché par les nombreux commentaires d'Howard, menant de manière tout à fait opportune à des éclaircissements dans mes propos.

Je souhaite adresser mes sincères remerciements envers Brest Métropole Océane pour son soutien financier, cette thèse n'aurait pu voir le jour sans cette aide.

Des échanges scientifiques que j'ai eu dans mon parcours scientifique, je retiens tout particulièrement les avis constructifs de Jean-Pierre Barthélemy, Sébastien Bigaret, Stéphane Lallich et Patrick Meyer. Qu'ils soient ainsi assurés de mes sentiments les meilleurs.

Je remercie tous les chercheurs, enseignants et membres du personnel du département Logique des Usages, Sciences Sociales et de l'Information (Enst Bretagne) et du département Statistique et Traitement Informatique des Données (Iut de Vannes) pour leur amitié et leur aide pendant mes travaux de thèse. Parmi toutes ces personnes, je remercie particulièrement Ghislaine, Sandrine et Emmanuelle.

J'adresse également mes remerciements à Carole et Janet pour leurs généreux travaux de relecture. Que ce temps passé ne soit pas ignoré.

Enfin, merci à toute ma famille pour son soutien constant. Et dans le désordre le plus total, merci à quelques amis qui ont compté ces dernières années : Thomas, Christophe et Fabienne, pis la petite Manon, Vivien, Thomas et Cathy, accompagnés par Circée, Amélie et Nicolas...

A ceux qui, chacun à leur façon, ont fait
ce bout de chemin à mes côtés, merci.

Benoît.

Table des matières

Introduction **1**

I L'Extraction de Connaissances à partir de Données **5**
1 Sur l'extraction de connaissances à partir de données . . . 6
2 Un modèle classique en ECD : la recherche de règles d'association . . . 7
2.1 Les règles d'association, définitions et notations . . . 7
2.2 L'algorithme AIS . . . 11

II Mesures de qualité de règles d'association **21**
1 Critère d'éligibilité des mesures de qualité . . . 22
2 Une "bonne" mesure de qualité : approche intuitive . . . 24
3 Mesures descriptives . . . 27
3.1 Support, confiance . . . 27
3.2 Coefficient de corrélation . . . 28
3.3 Coefficient de Loevinger . . . 28
3.4 Confiance centrée . . . 28
3.5 Conviction . . . 29
3.6 Estimateur de Laplace . . . 30
3.7 Facteur Bayésien . . . 30
3.8 Gain informationnel . . . 31
3.9 Indice de qualité de Cohen . . . 32
3.10 Lift . . . 32
3.11 Mesure de Fukuda . . . 32
3.12 Mesure de Ganascia . . . 33
3.13 Mesure de Jaccard . . . 34
3.14 Mesure de Sebag-Schoenauer . . . 34
3.15 Mesure de Piatetsky-Shapiro . . . 35
3.16 Moindre contradiction . . . 35
3.17 Taux d'exemples et de contre-exemples . . . 37
3.18 Zhang . . . 37
4 Indices statistiques et mesures probabilistes . . . 37
4.1 Indice d'implication . . . 39
4.2 Intensité d'implication . . . 39
4.3 Intensité d'implication entropique . . . 41
4.4 Intensité d'implication entropique tronquée et révisée . . . 42
4.5 Indice probabiliste discriminant . . . 44
4.6 Indice probabiliste d'écart à l'équilibre . . . 44

4.7 Indice probabiliste d'écart à l'équilibre entropique . . . 45
5 Synthèse . . . 46
6 Application aux données "jouet" . . . 46
7 Paramétrisation de mesures de qualité de règles d'association . . . 49
7.1 Mesures descriptives . . . 49
7.2 Mesures statistiques et probabilistes . . . 53
7.2.1 Rappels et complément . . . 53
7.2.2 Modélisations généralisées . . . 55
7.3 Pouvoir discriminant des mesures généralisées . . . 56

III Deux approches "aide multicritère à la décision" 63
1 Aspects théoriques . . . 66
1.1 Approche PROMETHEE-GAIA . . . 66
1.1.1 Comparaison paire par paire . . . 68
1.1.2 Indice de préférence agrégé . . . 69
1.1.3 Flots de surclassement . . . 69
1.1.4 Les rangements PROMETHEE . . . 70
1.1.5 Le plan GAIA . . . 70
1.2 Approche par critère unique de synthèse . . . 72
1.2.1 Présentation du processus d'agrégation . . . 74
1.2.2 Définition de relations valuées . . . 75
1.2.3 Propriétés des relations valuées . . . 75
1.2.4 Opérateurs d'agrégation . . . 78
2 Aide à la sélection de mesures de qualité . . . 80
2.1 Traitement orienté de la prémisse et de la conclusion . . . 82
2.2 Décroissance selon la taille du conséquent . . . 83
2.3 Evaluation de situations de références . . . 84
2.3.1 Règle logique . . . 84
2.3.2 Indépendance . . . 85
2.3.3 Indétermination . . . 85
2.3.4 Incompatibilité . . . 86
2.4 Tolérance aux premiers contre-exemples . . . 86
2.4.1 Modélisation classique . . . 86
2.4.2 Autres modélisations envisageables . . . 87
2.5 Prise en compte du nombre de cas . . . 90
2.6 Facilité à fixer un seuil d'acceptation de la règle . . . 90
2.6.1 Premières considérations . . . 91
2.6.2 Utilisation d'un paramètre de contrôle . . . 95
2.7 Intelligibilité de la mesure . . . 95
2.8 Analyse de l'espace des mesures . . . 95
2.9 Mise en œuvre : deux scenarii orientés utilisateur . . . 98
3 Agrégation de relations valuées . . . 101
3.1 Passage d'une mesure à une relation de préférence . . . 101
3.2 Choix d'opérateur d'agrégation . . . 105
3.3 Etude de cas . . . 105
3.3.1 Construction de relations de préférences floues . . . 106
3.3.2 Retour à une relation de préférence nette . . . 107

IV Expérimentations **121**
1 L'outil de simulation HERBS 121
1.1 Importation de données 123
1.2 Traitements mono-mesure 123
1.2.1 Sélection des N meilleures règles 126
1.2.2 Distribution des valeurs d'une mesure 127
1.3 Traitements multi-mesures 130
1.3.1 Distribution croisée des valeurs de deux mesures 130
1.3.2 Comparaison de préordres induits par deux mesures 132
1.3.3 Règles sélectionnées k fois parmi les meilleures 134
2 Expérimentations sur des bases de données réelles 135
2.1 Présentation des bases de données retenues 136
2.2 Présentation de résultats obtenus avec HERBS 139
3 Mise en regard des résultats théoriques et expérimentaux 139
4 Développements futurs 141

Conclusion et perspectives **143**

Bibliographie **145**

Summary **155**

Résumé **156**

Table des figures

I.1 Représentation schématique d'un processus d'ECD 6
I.2 Notations usuelles associées à une règle $\mathtt{A} \rightarrow \mathtt{B}$ 10
I.3 Treillis de Galois associé à un ensemble de 4 items $\{i_1, \ldots, i_4\}$ 12
I.4 Treillis de Galois associé à un ensemble de 5 items $\{i_1, \ldots, i_5\}$ 13
I.5 Treillis de Galois associé à un ensemble de 6 items $\{i_1, \ldots, i_6\}$ 14
I.6 Treillis de Galois associé à un ensemble de 7 items $\{i_1, \ldots, i_7\}$ 15

II.1 Variations possibles d'une mesure de qualité en fonction de $p_{a\bar{b}}$ 23
II.2 Ajout de contre-exemples à une règle, modélisation classique 24
II.3 Variations de SUP pour deux familles de règles 26
II.4 Variations de R en fonction des contre-exemples 28
II.5 Variations de LOE en fonction des contre-exemples 29
II.6 Variations de CONFCEN en fonction des contre-exemples 29
II.7 Variations de CONV en fonction des contre-exemples 30
II.8 Variations de FB en fonction des contre-exemples 31
II.9 Variations de GI en fonction des contre-exemples 31
II.10 Variations de IQC en fonction des contre-exemples 32
II.11 Variations de LIFT en fonction des contre-exemples 33
II.12 Variations de FUKUDA en fonction des contre-exemples ($\sigma_c = 0.6$) 33
II.13 Variations de GAN en fonction des contre-exemples 34
II.14 Variations de JAC en fonction des contre-exemples 34
II.15 Variations de SEB en fonction des contre-exemples 35
II.16 Variations de PS en fonction des contre-exemples 36
II.17 Variations de MOCO en fonction des contre-exemples 36
II.18 Variations de TEC en fonction des contre-exemples 37
II.19 Variations de ZHANG en fonction des contre-exemples 38
II.20 Variations de -INDIMP en fonction des contre-exemples 40
II.21 Variations de INTIMP en fonction des contre-exemples 40
II.22 Variations de INTIMP en fonction de n 41
II.23 Variations de IIE en fonction des contre-exemples 42
II.24 Variations croisées de IIE et RIIE, $n = 2000$ 43
II.25 Variations croisées de IIE et RIIE, $n = 200$ 43
II.26 Variations de IPEE en fonction des contre-exemples 45
II.27 Variations de IPEE en fonction de $p_{b|a}$ 46
II.28 Variations de $\widetilde{p}_x$ en fonction de p_x 59
II.29 Variations de mesures probabilistes et de leurs contreparties généralisées en fonction de $p_{b|a}$ 60

III.1 Problématique de choix . . . 65
III.2 Problématique de tri . . . 65
III.3 Problématique de rangement . . . 65
III.4 Problématique de synthèse . . . 65
III.5 Cinq situations de préférence construites sur $g_i(.)$ . . . 68
III.6 Fonction de préférence typique . . . 69
III.7 Fonction de préférence usuelle . . . 69
III.8 Comparaison d'une règle à fort conséquent avec une règle à faible conséquent . 83
III.9 Modélisation 1 . . . 88
III.10 Modélisation 2 . . . 88
III.11 Modélisation 3 . . . 89
III.12 Caractéristiques de $\mathtt{A}' \rightarrow \mathtt{B}'$ . . . 92
III.13 Dendrogramme associé à la CAH des mesures selon leurs propriétés formelles . 98
III.14 Stabilité des poids pour Sc1 . . . 100
III.15 Stabilité des poids pour Sc2 . . . 101
III.16 Plan Gaia associé à Sc1 . . . 102
III.17 Plan Gaia associé à Sc2 . . . 103
III.18 Fonction de préférence linéaire . . . 104
III.19 fonction de préférence gaussienne . . . 104
III.20 Graphe de préférences pour un seuil de coupure nul . . . 111
III.21 Graphe de préférences : σ_j pris à un écart de 40% . . . 112
III.22 Graphe de préférences : σ_j pris à un écart de 46% . . . 112
III.23 Graphe de préférences : σ_j pris à un écart de 53% . . . 112
III.24 Graphe de préférences : σ_j pris à un écart de 60% . . . 112
III.25 Graphe de préférences : σ_j pris à un écart de 66% . . . 113
III.26 Graphe de préférence pour $\lambda = 0.1$ . . . 114
III.27 Graphe de préférence pour $\lambda = 0.2$ . . . 115
III.28 Graphe de préférence pour $\lambda = 0.3$ . . . 116
III.29 Graphe de préférence pour $\lambda = 0.4$ . . . 116
III.30 Graphe de préférence pour $\lambda = 0.5$ . . . 116
III.31 Jeu de poids Sc1, σ_j pris à un écart de 5% . . . 117
III.32 Jeu de poids Sc1, σ_j pris à un écart de 25% . . . 117
III.33 Jeu de poids Sc1, σ_j pris à un écart de 50% . . . 118
III.34 Jeu de poids Sc1, σ_j pris à un écart de 75% . . . 118
III.35 Jeu de poids Sc2, $\lambda = 0,9$ . . . 118
III.36 Jeu de poids Sc2, $\lambda = 0,8$ . . . 118
III.37 Jeu de poids Sc2, $\lambda = 0,7$ . . . 119
III.38 Jeu de poids Sc2, $\lambda = 0,6$ . . . 119

IV.1 Architecture de l'outil Herbs . . . 122
IV.2 Importation d'une base de cas . . . 124
IV.3 Importation d'une base de règle . . . 124
IV.4 Choix des mesures à évaluer . . . 125
IV.5 Barre d'avancement des calculs . . . 125
IV.6 Menu contextuel associé à un couple de base de cas et de règles . . . 126
IV.7 Evolution de l'ensemble des N meilleures règles pour Conf, MoCo et TIIE . . 127
IV.8 Variations de 8 mesures en fonction de $n_{a\bar{b}}$ . . . 129

IV.9 Variations croisées de Conf et MoCo 131
IV.10 Variations croisées de Conf et TIIE 131
IV.11 Variations croisées de MoCo et TIIE 132
IV.12 Evolution de l'ensemble des N meilleures règles sélectionnées k fois 136

Liste des tableaux

I.1 Données "jouet" 8
I.2 Notations usuelles associées à une règle $\mathtt{A} \rightarrow \mathtt{B}$ 10
I.3 Règles extraites des données "jouet" 18

II.1 Valeurs de Sup pour deux familles de règles 26
II.2 Résumé des différentes modélisations 39
II.3 Mesures de qualité 47
II.4 Valeurs extrêmes des mesures 48
II.5 Evaluation des règles "jouet" par vingt mesures de qualité 50
II.6 Préordres induits sur les règles "jouet" par vingt mesures de qualité 51
II.7 Indices statistiques et mesures probabilistes 57
II.8 Indices statistiques et mesures probabilistes généralisés 57
II.9 Mesures descriptives paramétrisées, référence à l'indépendance 59
II.10 Mesures descriptives paramétrisées, référence à l'indétermination 60
II.11 Indices statistiques paramétrisés 60
II.12 Mesures probabilistes paramétrisées 61
II.13 Mesures probabilistes discriminantes paramétrisées 61

III.1 Conservation des propriétés par les opérateurs d'agrégation 81
III.2 Répartition des mesures pour les deux modalités de g_1 82
III.3 Répartition des mesures pour les deux modalités de g_2 84
III.4 Répartition des mesures selon les deux modalités de g_3 84
III.5 Répartition des mesures selon les deux modalités de g_4 85
III.6 Répartition des mesures selon g_3 et g_4 85
III.7 Répartition des mesures selon les deux modalités de g_5 86
III.8 Répartition des mesures selon les trois modalités de g_6 87
III.9 Table de contingence associée à la modélisation 1 88
III.10 Table de contingence associée à la modélisation 2 88
III.11 Table de contingence associée à la modélisation 3 89
III.12 Impact du choix de la modélisation sur l'évaluation de la propriété g_6 89
III.13 Répartition des mesures selon les deux modalités de g_7 90
III.14 Caractéristiques de $\mathtt{r}_1$ 91
III.15 Caractéristiques de $\mathtt{r}_2$ 91
III.16 Evaluation des règles $\mathtt{r}_1$ et $\mathtt{r}_2$ 91
III.17 Caractéristiques de $\mathtt{A}' \rightarrow \mathtt{B}'$ 92
III.18 Transformations linéaires de la confiance 93
III.19 Coefficients θ'_0 et θ'_1 93

III.20 Valeur maximale de α, cas des transformées linéaires . . . 93
III.21 Transformées monotones de CONF . . . 94
III.22 Valeurs maximales de α, cas des transformées monotones . . . 94
III.23 Variations de $\theta_{\max}$ en fonction de p_a, p_b et p_{ab} . . . 94
III.24 Répartition des mesures selon les trois modalités de g_9 . . . 95
III.25 Famille $\mathcal{F}$ de critères . . . 96
III.26 Matrice de synthèse sur les propriétés . . . 97
III.27 Matrice de synthèse des propriétés, sous forme normale disjonctive . . . 99
III.28 Rangements totaux associés aux scenarii SC1 et SC2 . . . 100
III.29 Différentes valeurs de σ . . . 107
III.30 Agrégation des préférences (exemple jouet, σ pris à 40%, poids équilibrés) . . . 108
III.31 Agrégation de préférences . . . 109
III.32 Jeu de poids utilisés dans les scénario SC1 et SC2 . . . 117

IV.1 Valeurs de INTIMP en fonction de $p_{a\bar{b}}$, avec $n_a = 0.25$ et $n_b = 0.4$ . . . 128
IV.2 Répartition des couples d'éléments de X^2 selon deux préordres $\mathcal{P}$ et $\mathcal{P}'$ . . . 134
IV.3 Présentation des données brutes . . . 137
IV.4 Présentation des données recodées . . . 137
IV.5 Présentation des données recodées . . . 137
IV.6 Présentation des bases de règles extraites par APRIORI . . . 138
IV.7 Comparaison des préordres induits par vingt mesures . . . 140
IV.8 Comparaison des classifications formelles et expérimentales . . . 140

Introduction

Lors du développement des premiers algorithmes d'apprentissage automatique en intelligence artificielle, conçus pour généraliser et constituer des connaissances à partir de données, ceux-ci étaient souvent confrontés à un volume relativement faible de données. De plus les données étaient généralement "simples". Depuis quelques années, la situation a radicalement changé.

Le développement de l'informatique et des moyens associés – en particulier, les capacités d'enregistrement et de calcul, la démocratisation des postes de travail individuels, etc. – permettent aujourd'hui de conserver une trace informatique de quasiment toutes les activités humaines et organisationnelles. Une première conséquence de ce développement est l'accroissement considérable du volume stocké et la diversité des données. Celles-ci sont désormais très hétérogènes (nombres, textes, images ou sons...), et il est couramment admis que leur volume double tous les dix-huit mois. Parallèlement on admet que l'analyse de ces mêmes données croît au plus de façon linéaire.

Pour autant, depuis l'apprentissage automatique, il est acquis qu'il existe des *structures* intéressantes – que l'on appellera désormais génériquement *connaissances*, ou *motifs* – à découvrir dans les données. Les besoins sont énormes. Tous les secteurs d'activité sont concernés, à des degrés divers : banque, assurance, finance, grande distribution, médecine, industrie, contrôle qualité, télécommunications, administration, etc. La découverte de structures intéresse toutes les fonctions et services des organisations, par exemple le marketing, les services financiers, le contrôle qualité, ou encore la sécurité. Toutes les connaissances produites par analyse et apprentissage sur les données peuvent alors servir à améliorer les décisions. Poussée par les besoins et les applications, favorisée par l'amélioration des moyens (capacités de calcul, développement des SGBDR, etc.), l'extraction de connaissances à partir de données émerge au début des années 1990, historiquement issue des travaux de nombreuses disciplines : l'apprentissage automatique, l'analyse de données, ou encore la statistique.

Parmi les nombreuses définitions de l'ECD nous en avons retenu deux :

- le *processus complexe permettant l'identification, au sein des données, de motifs valides, nouveaux, potentiellement intéressants et les plus compréhensibles possible* [Fayyad et al., 1996]
- le *processus interactif et itératif d'analyse d'un grand ensemble de données brutes afin d'en extraire des connaissances exploitables par l'utilisateur-analyste qui y joue un rôle central* [Kodratoff et al., 2001]

Chacune d'entre elles illustre différents points que nous avons abordés dans le travail pré-

senté dans ce mémoire. Elles insistent particulièrement toutes deux sur la notion de qualité et d'intérêt : les connaissances extraites doivent être valides, intéressantes, exploitables et compréhensibles... par un utilisateur final. La notion d'intérêt peut donc varier en fonction du contexte applicatif et de l'utilisateur final.

Une autre caractéristique du domaine de l'ECD est le traitement de volumes importants de données. De gros progrès ont été réalisés depuis le début et les algorithmes ou méthodes de traitement de données sont chaque jour de plus en plus performants. Associés aux capacités croissantes des systèmes informatiques, ces progrès permettent de traiter des volumes très importants de données. Une conséquence directe est alors que le volume des connaissances extraites est lui aussi colossal. L'utilisateur final n'est donc plus en mesure de valider les connaissances sans assistance.

La validation de ces connaissances est ainsi une étape cruciale du processus [Hilderman and Hamilton, 2001b]. Elle devient particulièrement incontournable lors de l'évaluation de règles d'association sur lesquelles nos recherches ont porté. Les algorithmes de la famille APRIORI [Agrawal and Srikant, 1994], qui déterminent de telles règles, peuvent en effet rapidement fournir un très grand nombre de motifs. Il existe différentes manières d'aborder ce sujet. Parmi celles-ci, les – nombreuses – mesures objectives de qualité servent à filtrer, classer, ordonner les règles selon leur intérêt : chaque mesure évalue quantitativement l'intérêt d'une règle. Cependant, les mesures d'intérêt peuvent classer très différemment les règles.

Partant du constat que la sélection des *bonnes* connaissances passe aussi par l'utilisation des *bonnes* mesures [Tan et al., 2002, Lenca et al., 2003a, Carvalho et al., 2005], nous présentons une étude systématique des mesures d'intérêt des règles d'association selon différents axes d'analyse. Notre approche est volontairement indépendante du contexte applicatif et de connaissances *a priori* sur le domaine. Elle est ainsi *générique*.

Le mémoire est divisé en quatre grands chapitres.

Après un bref résumé des différentes étapes du processus d'ECD, le premier chapitre présente particulièrement les règles d'association (*type de connaissances* sur lequel porte notre travail) et la famille des algorithmes de type APRIORI qui permet de déterminer de telles règles. Nous présentons brièvement différentes solutions proposées pour y faire face. Parmi elles, les mesures de qualité peuvent jouer un rôle majeur, en complément des autres solutions.

Le second chapitre présente le critère d'éligibilité que nous retenons pour les mesures d'intérêt des règles d'association. Sans dévoiler complètement le troisième chapitre on présentera intuitivement les qualités attendues d'une *bonne mesure* de qualité. On présentera alors les principales mesures descriptives, les indices statistiques et les mesures probabilistes définis dans la littérature et étudiés dans ce mémoire. On propose enfin un cadre permettant de généraliser et paramétrer un grand nombre de mesures.

Le troisième chapitre présente deux analyses d'aide à la décision. La première cherche à fournir une aide multicritère à la décision pour l'utilisateur. Le but recherché consiste à recommander l'utilisation de bonnes mesures en fonction des besoins de ce dernier. La seconde cherche à tenir compte de l'avis de différentes mesures pour l'évaluation d'une règle, c'est-à-dire agréger les valeurs prises par les mesures. Nous présentons dans un premier temps les éléments nécessaires à la compréhension des deux approches. Ce chapitre sera

l'occasion de présenter une étude systématique de propriétés *souhaitables* pour les mesures et les opérateurs d'agrégation. Nous y présentons une classification *formelle* et les résultats des deux analyses d'aide à la décision seront présentés pour vingt mesures.

Le quatrième chapitre présente une étude en situation de mesures listées au chapitre II. Pour ce faire nous avons développé un outil informatique performant, HERBS, permettant de systématiser des études expérimentales sur des données. Système ouvert, il permet d'intégrer facilement de nouvelles mesures. Il permet de mener des études mono-mesures (comme le calcul de l'ensemble des `N` meilleures règles pour une mesure donnée) et multi-mesures (comme la distribution croisée des valeurs de deux mesures). Nous proposons alors une classification *expérimentale* des mesures. Cette dernière est confrontée à la classification *formelle* du chapitre III.

Chaque chapitre est illustré selon les besoins à partir de jeux de données *jouet* ou bien *réels*.

Nous conclurons ensuite notre travail et présenterons quelques perspectives de celui-ci.

Chapitre I

L'Extraction de Connaissances à partir de Données

Les définitions de l'ECD sont nombreuses. Nous en avons retenu deux, illustrant le contexte dans lequel nous nous positionnons : il s'agit du *processus complexe permettant l'identification, au sein des données, de motifs valides, nouveaux, potentiellement intéressants et les plus compréhensibles possible* [Fayyad et al., 1996], du *processus interactif et itératif d'analyse d'un grand ensemble de données brutes afin d'en extraire des connaissances exploitables par l'utilisateur-analyste qui y joue un rôle central* [Kodratoff et al., 2001].

Ces deux définitions illustrent chacune différents points que nous allons aborder par la suite. Au centre d'un processus non trivial, se situe un ou des utilisateurs, dont l'objectif est d'exploiter un ensemble de données. L'accent est mis sur la notion de qualité et d'intérêt. Les connaissances que l'on peut découvrir lors du déroulement de ce processus doivent être valides, intéressantes, exploitables et compréhensibles... par un utilisateur final. La notion d'intérêt prend donc son sens au sein d'un contexte applicatif et de l'utilisateur final.

L'ECD s'intéresse ainsi à des problématiques d'analyse de données. Elle se distingue toutefois des approches classiques, car plus qu'une méthodologie, il s'agit d'un cadre d'étude ancré dans un contexte applicatif. En effet, elle ne traite pas d'une solution permettant de résoudre les problèmes abordés, mais d'un ensemble de concepts, à moduler en fonction des caractéristiques de la problématique abordée.

Parmi les applications classiques, on trouvera par exemple la recherche prédictive, la logique implicative, ou encore la classification (clustering).

On cite souvent en exemple les applications médicales, où l'on souhaite, détecter une ou des maladies que pourrait avoir un patient, à partir d'un ensemble de relevés cliniques. Le domaine bancaire est également classique : il est très intéressant de distinguer les bons des mauvais emprunteurs lorsque l'on accorde un prêt.

Nous présentons brièvement dans ce chapitre les principales étapes de l'ECD, et explorons un cadre d'étude particulier, celui de la recherche de règles d'association. Les travaux présentés dans ce mémoire sont extensibles, mutatis mutandis, aux mesures de qualité issues

d'autres contextes de fouille de données.

1 Sur l'extraction de connaissances à partir de données

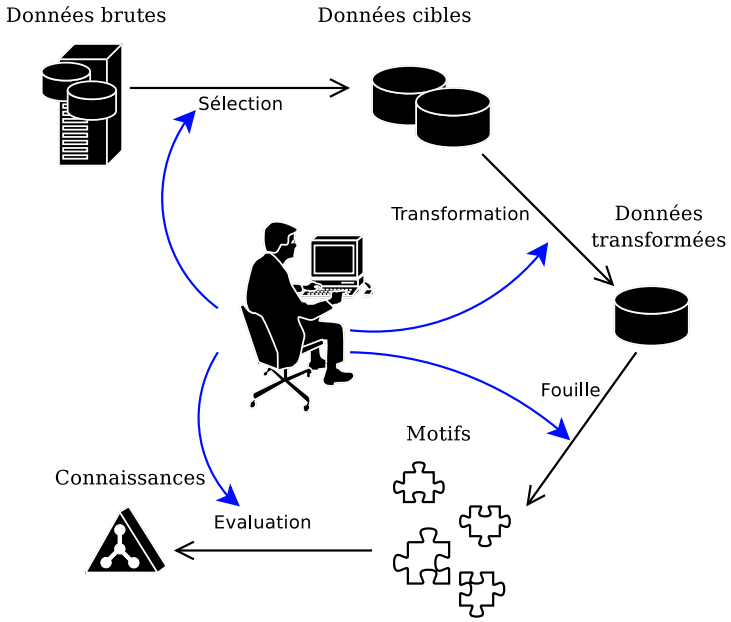


FIG. I.1 – Représentation schématique d'un processus d'ECD

De manière usuelle, on décompose un processus d'ECD en plusieurs étapes (voir figure I.1). Chaque étape, spécifique au problème traité, est validée par un utilisateur (tantôt un expert métier, tantôt un expert de l'ECD). L'orientation du processus, et donc les connaissances qui en seront extraites, se fait principalement lors de la troisième étape, celle de fouille de données. En effet, la type de motifs extraits dépend fortement du choix de l'algorithme de fouille retenu. Les étapes majeures d'un processus d'ECD sont ainsi :

La sélection Toutes les informations à disposition n'étant pas nécessairement pertinentes pour une application donnée (par exemple, disposer des données relatives à tous les clients d'une banque lorsque l'on souhaite extraire des informations relatives aux "jeunes" semble inutile), il est nécessaire de sélectionner une sous-partie adaptée à l'étude à mener.

La transformation En vue d'appliquer un traitement informatique aux informations précédemment sélectionnées, il peut être nécessaire d'adapter leur codage à la méthode de fouille retenue pour la troisième étape. En effet, certaines méthodes s'appliqueront à des données constituées de variables discrètes, voire uniquement binaires, d'autres pourront s'appliquer à des variables continues. Les résultats de l'étape de fouille ne sauraient être garantis comme intéressants si les données en entrée de la méthode appliquée ne sont pas en adéquation avec ce qu'elle peut traiter.

La fouille L'étape de fouille est au cœur du processus d'ECD car c'est lors de son déroulement qu'à partir des données, *a priori* inexploitables en l'état, on en extrait des motifs

à partir des informations qu'elles contiennent. Selon la méthode de fouille choisie, on pourra extraire différents types de motifs (nous détaillerons un modèle particulier en section 2), permettant de focaliser sur des problématiques particulières, comme la logique prédictive (dans le cas des arbres de décision), l'analyse exploratoire (dans le cas des règles d'association), ou encore la synthèse de bases de données (dans le cas des résumés de bases de données).

L'évaluation Les résultats produits par les algorithmes de fouille de données ne sont pas toujours exploitables directement. En effet, il se peut qu'un important volume de motifs soient extraits des données, ces motifs relevant éventuellement de phénomènes statistiques, des niveaux de bruit ou des erreurs contenues dans les données initiales. Ainsi, il est utile d'automatiser un filtrage ou un ordonnancement automatique de ces motifs, afin d'assister un utilisateur expert des données dans la tâche de prise de décision, ou d'interprétation des résultats obtenus. Ceci peut se faire au moyen de mesures de qualité, fonctions mathématiques modélisant une catégorie de connaissances désirée, et mettant celle-ci en avant parmi l'ensemble des motifs extraits. L'étude comparative de mesures de qualité constitue le cœur de ce mémoire.

2 Un modèle classique en ECD : la recherche de règles d'association

Nous présentons dans cette section un paradigme classique en fouille de données, celui de l'extraction de règles d'associations. On s'intéresse à exploiter un ensemble de données décrivant des individus. Ceux-ci sont décrits par des attributs (leurs caractéristiques connues). Ce paradigme est dit "non supervisé", au sens où l'on ne cherchera pas à prédire un attribut particulier, mais à effectuer une recherche tous azimuts dans les données, afin de détecter des liens de causalité entre différents attributs. Par exemple, on cherchera à analyser quels produits achetés par un client permettent d'inférer que d'autres le seront également par ce même client.

A l'inverse, en apprentissage "supervisé", on cherchera à déduire un attribut particulier, dit attribut cible, qui est généralement la classe à laquelle appartiennent les objets étudiés, à partir d'un ensemble d'attributs prédictifs. Ainsi, à partir de relevés cliniques, on pourra chercher à prédire l'état du patient (sain ou malade), à partir d'indicateurs financiers, ou pourra souhaiter savoir si une banque devrait ou non accorder un prêt à l'un de ses clients...

2.1 Les règles d'association, définitions et notations

On suppose que les données D (un ensemble d'objets, ou de transactions) à explorer sont binaires, *i.e.* qu'on peut décrire chaque transaction au moyen d'un ensemble fini d'attributs $\mathcal{I} = \{i_1, \ldots, i_m\}$, également appelés *items*. Chaque transaction t sera donc un sous-ensemble de $\mathcal{I}$. De plus, on associe à chaque transaction un identifiant (TID pour *"Transaction IDentifier"*) : $D = \{t_1, \ldots, t_n\}$. La table I.1 illustre ce type de données, pour un ensemble de 10 transactions $\{t_1, \ldots, t_{10}\}$ décrit par 7 items $\{i_1, \ldots, i_7\}$. Nous utiliserons ce jeu de données volontairement minimal afin d'illustrer au besoin les notions introduites dans la suite de ce mémoire.

Pour un ensemble X d'items (on parle généralement d'itemsets) de $\mathcal{I}$, on dira qu'une transaction t contient X si et seulement si $X \subseteq t$. Par exemple, l'itemset $\{i_3, i_4\}$ est contenu dans la transaction t_1.

t_1	i_3	i_4	i_5			
t_2	i_1	i_3	i_5	i_6		
t_3	i_1	i_3	i_4	i_7		
t_4	i_2	i_4	i_6			
t_5	i_1	i_2	i_4	i_5	i_7	
t_6	i_1	i_2	i_3	i_4	i_5	i_7
t_7	i_2	i_4	i_5	i_6		
t_8	i_1	i_4	i_6			
t_9	i_3	i_4	i_6			
t_{10}	i_2	i_3	i_5	i_6		

TAB. I.1 – Données "jouet"

Définition 1 (support) *On appelle support d'un itemset X le rapport entre le nombre de transactions contenant X et le nombre total de transactions de D. Cette proportion sera notée* SUP.

$$\text{SUP}(X) = \frac{|\{t \in D : X \subseteq t\}|}{|D|}$$

En d'autres termes, le support d'un itemset correspond à la fréquence d'apparition de celui-ci dans les données. Dans les données "jouet" listées précédemment, on a par exemple :

$$\text{SUP}(\{i_1\}) = \frac{|\{t_2, t_3, t_5, t_6, t_8\}|}{10} = 5/10 = 50\%$$

$$\text{SUP}(\{i_3, i_4\}) = \frac{|\{t_1, t_3, t_6, t_9\}|}{10} = 4/10 = 40\%$$

De par la définition d'un itemset, on a la propriété suivante :

Propriété 1 (antimonotonicité) *Etant donné deux itemsets X et Y, on a :*

$$X \subseteq Y \Rightarrow \text{SUP}(X) \geq \text{SUP}(Y)$$

Preuve 1 *La preuve de cette propriété est triviale, car toutes les transactions contenant Y contiennent également X.*

Cette propriété, connue en ECD sous le nom de propriété d'*antimonotonicité* du support, est cruciale pour l'algorithme AIS que nous développerons en section 2.2.

Propriété 2 *Moins intuitivement, notons aussi que :*

$$\text{SUP}(X \cup Y) \leq \min(\text{SUP}(X), \text{SUP}(Y)) \leq \max(\text{SUP}(X), \text{SUP}(Y)) \leq \text{SUP}(X \cap Y)$$

Preuve 2 *En effet, plus un itemset est grand, plus le nombre de transactions le contenant est faible, d'où une diminution de son support. L'union de deux itemsets est l'intersection des transactions les contenant.*

On peut aisément vérifier cela sur notre exemple "jouet" :

$$\text{SUP}(\{i_1, i_3, i_4\}) = \frac{|\{t_3, t_6\}|}{10} = 2/10 = 20\% \leq \min(\text{SUP}(\{i_1\}) = 50\%, \text{SUP}(\{i_3, i_4\}) = 40\%)$$

Définition 2 (itemset fréquent) *On dit qu'un itemset X est fréquent lorsque son support est supérieur ou égal à un seuil σ donné. Un itemset n'ayant pas le support suffisant est dit non fréquent.*

Ainsi, en choisissant un seuil $\sigma = 30\%$, on obtient sur les données "jouet" que $\{i_1\}$ et $\{i_3, i_4\}$ sont fréquents, mais que $\{i_1, i_3, i_4\}$ ne l'est pas. En revanche, en choisissant un seuil à 10%, les trois itemsets sont fréquents.

Propriété 3 *Etant donné un seuil de référence σ fixé,*
- *tout sous-ensemble d'un itemset fréquent est fréquent,*
- *tout sur-ensemble d'un itemset non fréquent est non fréquent.*

Preuve 3 *Soit σ un seuil de référence donné. Tout itemset de support supérieur ou égal à σ est fréquent.*

Soit X' un sous-ensemble d'un itemset fréquent X. D'après la propriété 1, on a :

$$X' \subseteq X \Rightarrow \text{SUP}(X') \geq \text{SUP}(X)$$

X étant fréquent, $\text{SUP}(X) \geq \sigma$. Ainsi, $\text{SUP}(X') \geq \sigma$, et X' est donc fréquent.

De par cette même propriété, Y' étant un sur-ensemble d'un itemset Y non fréquent, on déduit que :

$$Y \subseteq Y' \Rightarrow \text{SUP}(Y) \geq \text{SUP}(Y')$$

Y étant non fréquent, $\sigma > \text{SUP}(Y)$. Il vient alors $\sigma > \text{SUP}(Y')$, et Y' est donc non-fréquent.

Définition 3 (Règle d'association, support, confiance) *Une règle d'association est un couple (A,B), où A et B sont des itemsets non vides disjoints, i.e. $\mathtt{A} \neq \emptyset$, $\mathtt{B} \neq \emptyset$, et $\mathtt{A} \cap \mathtt{B} = \emptyset$. On note classiquement un tel couple sous la forme $\mathtt{A} \rightarrow \mathtt{B}$.*

On définit le support d'une règle d'association comme étant le support de l'itemset $\mathtt{A} \cup \mathtt{B}$ (i.e. la proportion de transactions contenant à la fois A et B) :

$$\text{SUP}(\mathtt{A} \rightarrow \mathtt{B}) = \text{SUP}(\mathtt{A} \cup \mathtt{B}) = \frac{|\{t \in D : \mathtt{A} \subseteq t \text{ et } \mathtt{B} \subseteq t\}|}{|D|}$$

On définit également la confiance d'une règle $\mathtt{A} \rightarrow \mathtt{B}$, notée $\text{CONF}(\mathtt{A} \rightarrow \mathtt{B})$, comme étant le rapport entre le support de l'itemset $\mathtt{A} \cup \mathtt{B}$ et de celui de A :

$$\text{CONF}(\mathtt{A} \rightarrow \mathtt{B}) = \frac{\text{SUP}(\mathtt{A} \cup \mathtt{B})}{\text{SUP}(\mathtt{A})}$$

A condition d'approximer la probabilité p_x d'observer un évènement X par sa fréquence $f_x = \text{SUP}(X)$ dans D, le support et la confiance d'une règle d'association $\texttt{A} \rightarrow \texttt{B}$ se lisent alors respectivement comme la probabilité conjointe d'observer les évènements A et B, et la probabilité conditionnelle d'observer B, sachant A.

Etant donné une règle $\texttt{A} \rightarrow \texttt{B}$, il est ainsi courant de l'analyser en se rapportant à une matrice de contingence croisant deux variables binaires A et B. On obtient alors une description de notre règle sous l'une des formes listées dans le tableau I.2. n_x correspond au nombre de transactions contenant X (ou fréquence absolue), et p_x à la probabilité de X (ou fréquence relative).

La figure I.2 illustre ces notations en fréquences absolues. Par convention, on note multiplicativement l'union de deux itemsets afin d'alléger l'écriture (ainsi, le nombre de transactions contenant $X \cup Y$ est n_{xy}). La première table présente ces notations en fréquences absolues, la seconde en fréquences relatives.

Ces notations sont équivalentes, on passe de l'une à l'autre simplement via la relation $p_x = n_x/n$, avec $n = |D|$. Par la suite, on utilisera selon les besoins la notation la mieux adaptée.

A marges fixées, il suffit de connaître l'une des cellules du tableau de contingence pour en déduire les autres. En effet, ce tableau n'a que trois degrés de liberté. Les cellules du tableau de contingence sont liées par les relations suivantes :

- $p_a + p_{\bar{a}} = 1$
- $p_b + p_{\bar{b}} = 1$
- $p_{ab} = p_b - p_{\bar{a}b} = p_a - p_{a\bar{b}}$
- $p_{a\bar{b}} = p_a - p_{ab} = p_{\bar{b}} - p_{\bar{a}\bar{b}}$
- $p_{\bar{a}b} = p_b - p_{ab} = p_{\bar{a}} - p_{\bar{a}\bar{b}}$
- $p_{\bar{a}\bar{b}} = p_{\bar{b}} - p_{a\bar{b}} = p_{\bar{a}} - p_{\bar{a}b}$

$A \backslash B$	0	1	total
0	$n_{\bar{a}\bar{b}}$	$n_{\bar{a}b}$	$n_{\bar{a}}$
1	$n_{a\bar{b}}$	n_{ab}	n_a
total	$n_{\bar{b}}$	n_b	n

$A \backslash B$	0	1	total
0	$p_{\bar{a}\bar{b}}$	$p_{\bar{a}b}$	$p_{\bar{a}}$
1	$p_{a\bar{b}}$	p_{ab}	p_a
total	$p_{\bar{b}}$	p_b	1

TAB. I.2 – Notations usuelles associées à une règle $\texttt{A} \rightarrow \texttt{B}$

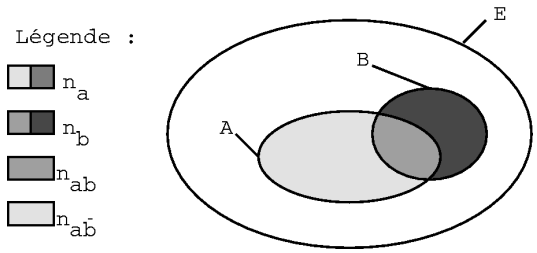


FIG. I.2 – Notations usuelles associées à une règle $\texttt{A} \rightarrow \texttt{B}$

Sur ces bases classiques en statistique, il viendra naturellement de considérer diverses situations de référence afin d'évaluer une règle, telle la règle logique (dans ce cas, la règle

$\mathtt{A} \rightarrow \mathtt{B}$ n'a pas de contre-exemples, *i.e.* lorsque $n_{a\bar{b}} = 0$), l'indépendance (*i.e.* lorsque $n_{ab} = n_a n_b / n$), l'indétermination (*i.e.* lorsque $n_{ab} = n_a/2$), ou encore l'incomparabilité (*i.e.* lorsque $n_{ab} = 0$). Ces notions seront détaillées en section 2.3 du chapitre III.

2.2 L'algorithme AIS

L'objectif de l'algorithme AIS [Agrawal et al., 1993] est, étant donné un jeu de données D, d'en extraire toutes les règles d'association vérifiant des seuils minimaux de support et de confiance donnés. Cet algorithme s'appuie sur une structure de donnée bien particulière : le treillis de concept. Sans plus entrer dans les détails de cette structure, notons toutefois qu'elle va permettre une exploration exhaustive des données, sous certaines contraintes définies par l'utilisateur.

Afin de procéder à l'extraction, les auteurs proposent de procéder en deux phases. Ils se restreignent également à l'extraction de règles dont la conclusion n'est formée que d'un singleton. Notant σ_s et σ_c les seuils respectivement de support et de confiance, on procède comme suit :

1. Générer l'ensemble de tous les itemsets fréquents, *i.e.* de support supérieur ou égal à σ_s,
2. Pour chaque itemset fréquent X, considérer les règles de la forme $X - Y \rightarrow Y$, où Y est un singleton de X, et ne retenir que celles ayant une confiance supérieure ou égale à σ_c.

Pour chaque itemset fréquent de cardinalité k (ou k-itemset), il y a donc au plus k règles distinctes, la conclusion étant par hypothèse constituée d'un seul item.

La difficulté majeure est d'obtenir la liste de tous les itemsets fréquents, en particulier en présence de données volumineuses. Les auteurs utilisent la propriété d'antimonotonicité du support afin d'élaguer l'espace de recherche des itemsets fréquents.

L'algorithme 1, proposé par [Agrawal et al., 1993] génère tous les itemsets fréquents en au plus n passes sur les données, n étant le nombre d'items. Cet algorithme fonctionne sur un principe incrémental : chaque passe sur les données fait évoluer une frontière contenant des itemsets aux cardinalités croissantes, jusqu'à épuisement des sur-itemsets fréquents dans le treillis. A la k^{e} passe, la frontière ne contiendra par construction que des itemsets de cardinalité supérieure ou égale à k. On va donc faire progresser de proche en proche dans le treillis des itemsets une frontière constituée d'itemsets fréquents. Très récemment, [Szathmary et al., 2006] se sont au contraire intéressés à l'extraction de motifs rares.

Pour des raisons de clarté, deux fonctions sont étudiées séparément : la fonction $IC(.,.)$ qui, étant donné un itemset fréquent et une transaction, retourne un ensemble de sur-itemsets potentiellement fréquents (appelés "itemsets candidats" dans l'algorithme), et la fonction $frontiere(.)$ qui est vraie si et seulement si l'itemset en argument doit être considéré à la passe suivante sur la base de données.

Les figures I.3 à I.6 présentent les treillis de Galois associés à des ensembles de 4, 5, 6 et 7 items (ce dernier cas correspondant à celui des données du tableau I.1).

Clairement, il existe généralement plusieurs chemins pour arriver en un nœud du treillis en partant de l'ensemble vide. Afin de ne pas considérer plusieurs fois un même nœud lors

du parcours du treillis, [Agrawal et al., 1993] se munissent d'un ordre $\succ$ sur les items (tout simplement l'ordre lexicographique, mais n'importe quel ordre total est admissible). Ainsi, on garantit que l'algorithme 1 passera une fois au plus par chaque nœud fréquent du treillis.

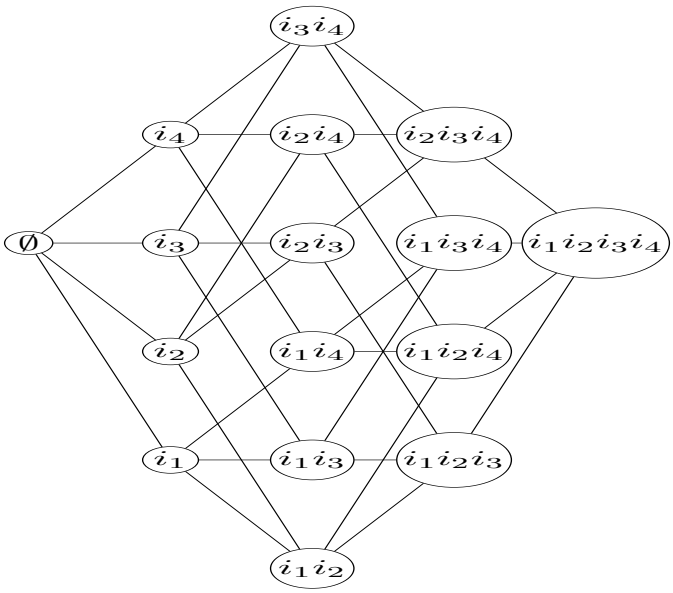


FIG. I.3 – Treillis de Galois associé à un ensemble de 4 items $\{i_1, \ldots, i_4\}$

La fonction $IC(X, t)$ prend comme arguments un itemset fréquent X et un tuple t. Elle retourne un ensemble d'itemsets potentiellement fréquents. On pourrait naïvement penser retourner l'ensemble des itemsets obtenus en ajoutant un item de t à X, pourvu que celui-ci n'y soit pas déjà inclus. Toutefois, cette stratégie est mauvaise pour deux raisons.

D'une part, et afin d'optimiser le parcours du treillis, on ne considérera pas tous les items de t afin d'augmenter X. On se restreindra aux items de t supérieurs au plus grand des items de X (selon l'ordre $\succ$ que l'on s'est donné sur les itemsets). Ceci se fait sans risque puisque s'il existe un itemset X' fréquent, contenant X et des items de t, non inclus dans X et inférieurs au plus grand des éléments de X au sens de l'ordre sur $\mathcal{I}$, alors cet itemset sera construit lorsque l'on étendra l'itemset $X \cap t$ ou ses sur-itemsets, nécessairement tous fréquents puisque sous-itemsets de X', que l'on suppose fréquent par hypothèse.

Posant $l = \max\{j \in X\}$, on considérera donc l'ensemble $I = \{x \in t | x \succ l\}$ afin de générer des sur-itemsets de X. Minimalement, $IC(X, t)$ retournera donc l'ensemble d'itemsets $\{\{i\} \cup X, i \in I\}$ (éventuellement vide dans le cas où I l'est).

D'autre part, ne retourner que cet ensemble peut ne pas être optimal. Supposons $X = \{\alpha, \beta\}$ et que $t = \{\alpha, \beta, \gamma, \delta\}$, l'ordre sur les items étant l'ordre alphabétique. On a $I = \{\gamma, \delta\}$, et au minimum $\{\{\alpha, \beta, \gamma\}, \{\alpha, \beta, \delta\}\} \subseteq IC(X, t)$. Si $\{\alpha, \beta, \gamma, \delta\}$ est fréquent, et que seuls les itemsets d'une unité supérieure sont retournés par $IC(.,.)$, il sera nécessaire d'effectuer deux passes sur nos données afin d'exhiber l'itemset fréquent $\{\alpha, \beta, \gamma, \delta\}$. Ainsi, on formalise la notion de support attendu comme suit.

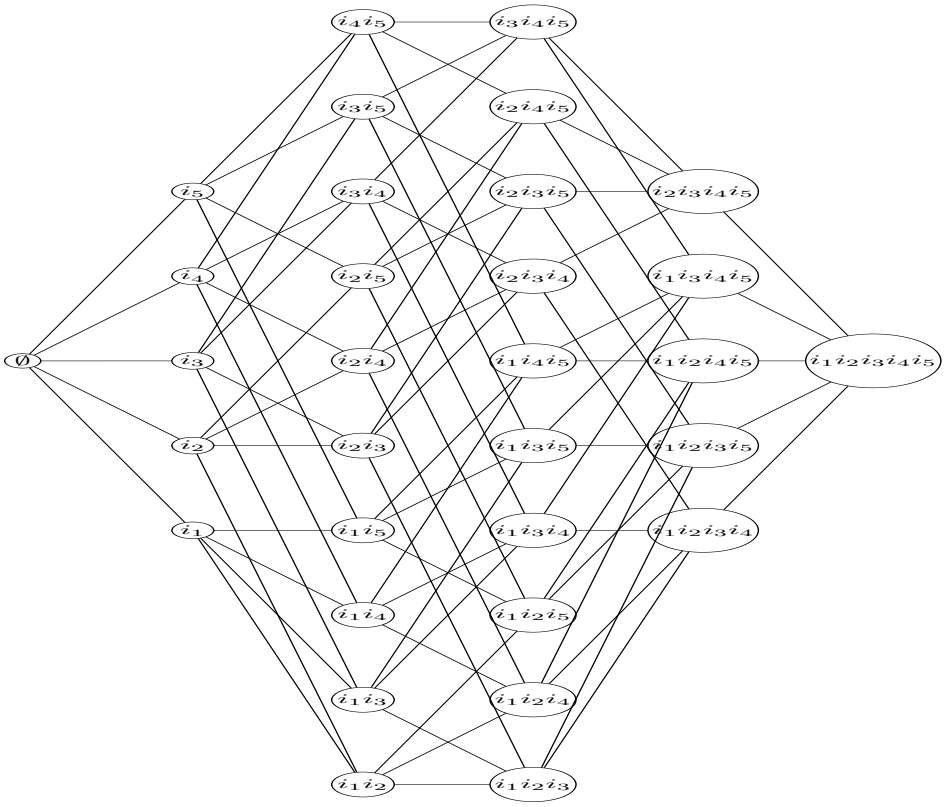


FIG. I.4 – Treillis de Galois associé à un ensemble de 5 items $\{i_1, \ldots, i_5\}$

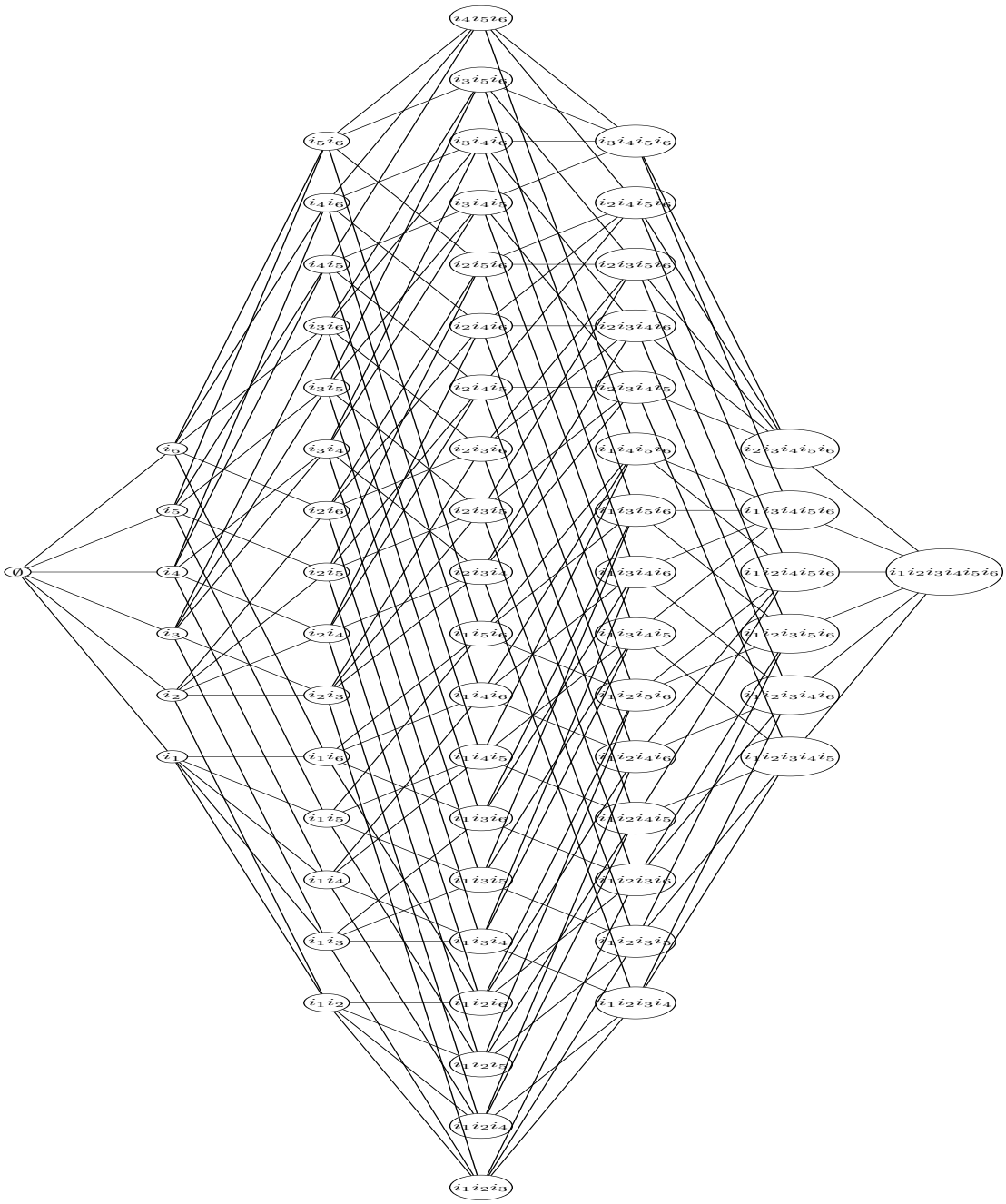


FIG. I.5 – Treillis de Galois associé à un ensemble de 6 items $\{i_1, \ldots, i_6\}$

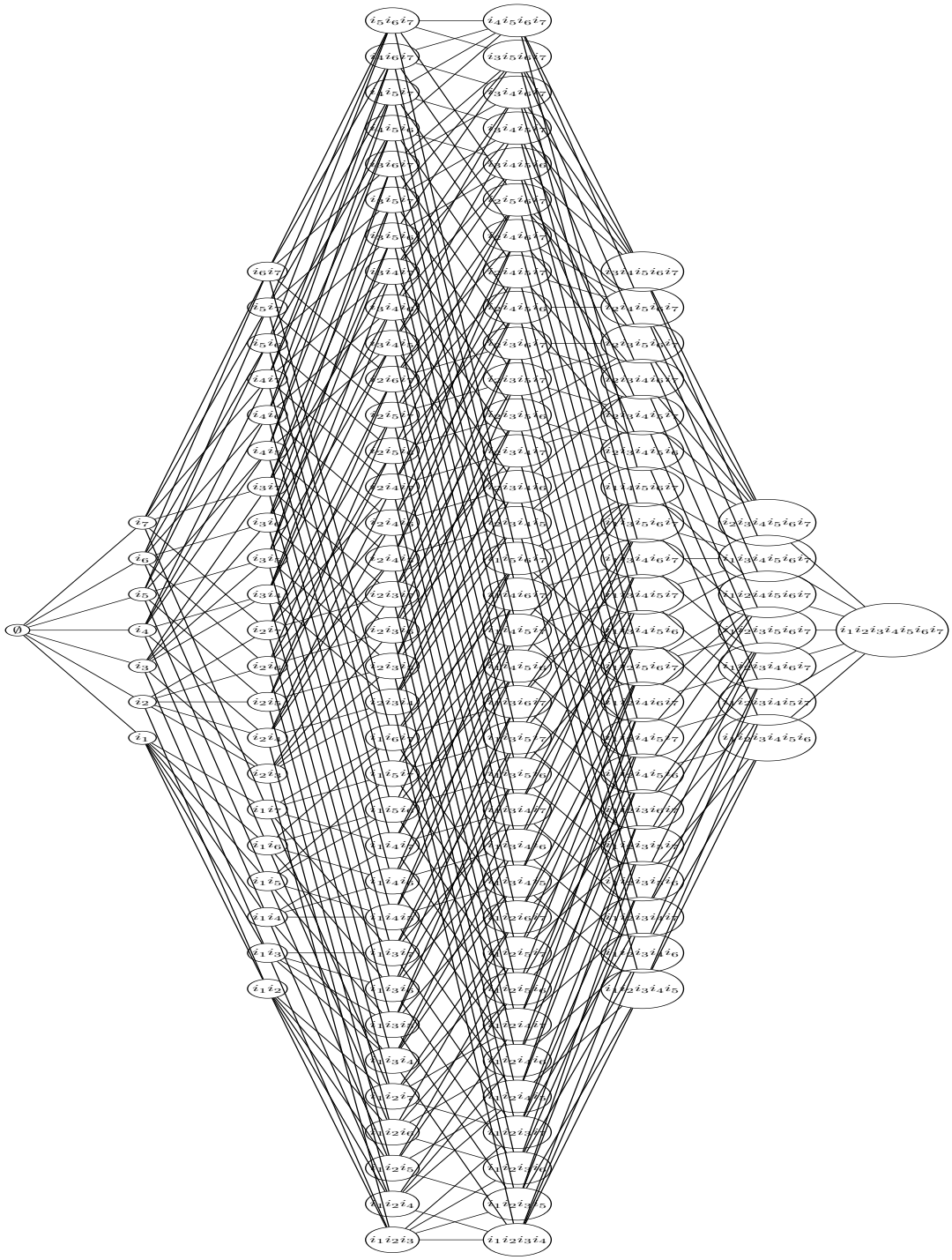


FIG. I.6 – Treillis de Galois associé à un ensemble de 7 items $\{i_1, \ldots, i_7\}$

Algorithme 1: Algorithme de génération des itemsets fréquents

Données : Données d'apprentissage D, seuil de support minimal σ
Résultat : Générer tous les itemsets fréquents L
$L \leftarrow \emptyset$ # initialisation de l'ensemble des itemsets fréquents
$F \leftarrow \{\emptyset\}$ # initialisation de la frontière
tant que $F \neq \emptyset$ **faire**
  # faire une passe sur les données pour générer les candidats
  $C \leftarrow \emptyset$
  pour tous les $t \in D$ **faire**
    pour tous les $f \in F$ **faire**
      si $f \in t$ **alors**
        $C_f = IC(f, t)$
        pour tous les $c_f \in C_f$ **faire**
          si $c_f \in C$ **alors**
            $c_f.count \leftarrow c_f.count + 1$
          sinon
            $c_f.count \leftarrow 0$
            $C \leftarrow C \cup \{c_f\}$
          fin
        fin
      fin
    fin
  fin
  # mettre à jour la frontière
  $F \leftarrow \emptyset$
  pour tous les $c \in C$ **faire**
    si $\text{SUP}(c) = c.count/|D| \geq \sigma$ **alors**
      $L \leftarrow L \cup \{c\}$
    fin
    si $frontiere(c)$ **alors**
      $F \leftarrow F \cup \{c\}$
    fin
  fin
fin

Définition 4 (support attendu) *Etant donné un itemset $X \cup Y$, extension de l'itemset X appartenant à la frontière lors de la passe courante de l'algorithme* AIS*, on appelle support attendu de $X \cup Y$ la grandeur :*

$$\bar{s}_t(X \cup Y) = \prod_{j=1}^{n} f(Y[j]) \times (f(X) - \frac{X.count}{|D|})$$

où $f(.)$ correspond à la fréquence d'occurrence de l'itemset dans D, avec $f(\emptyset) = 1$, $Y[j] = \{i_j\}$ si $i_j \in Y$, $\emptyset$ sinon, et $X.count$ est le nombre de tuples contenant X déjà observés lors de la passe courante.

On dira de plus que l'itemset $X \cup Y$ a un support attendu fréquent lorsque $\bar{s}_t(X \cup Y) \geq \sigma_s$. Inversement, si $\bar{s}_t(X \cup Y) < \sigma_s$, on dira qu'il a un support attendu non fréquent.

Remarques : $X \cup Y$ étant extension de X, on a :

- $X \cap Y = \emptyset$,
- de plus, $\forall y \in Y, y \succ \max\{x \in X\}$,
- la grandeur $f(X)$ est connue puisque calculée lors de la passe précédente sur les données,
- afin de générer $IC(X,t)$, il suffit de procéder par récurrence (voir algorithme 2).

Algorithme 2: Algorithme de génération des itemsets candidats : procédure $IC(.,.)$

Données : Itemset $X \cup Y$, tuple t
Résultat : Générer l'ensemble d'itemsets candidats associés à $X \cup Y$ et t
$I \leftarrow \{y \in t | y \succ \max\{x \in X\}\}$ # ensemble des items de t à considérer
$C \leftarrow \emptyset$ # ensemble des candidats
pour tous les $i \in I$ **faire**
  $C \leftarrow C \cup \{X \cup Y \cup \{i\}\}$
  si $\bar{s}_t(X \cup Y \cup i) < \sigma$ **alors**
    $C \leftarrow C \cup IC(X \cup Y \cup \{i\}, t)$ # appel récursif
  fin
fin
retourner C

Enfin, afin de construire l'ensemble frontière en fin de passe de l'algorithme AIS, il suffit de ne considérer que ceux ayant été prédits comme attendus non fréquents mais qui se sont révélés fréquents.

L'un des outils les plus utilisés permettant d'extraire des règles d'association est l'implémentation d'APRIORI, une version améliorée d'AIS, proposée dans [Agrawal and Srikant, 1994]. L'implémentation a été proposée par [Borgelt and Kruse, 2002], et disponible à l'adresse `http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori.html`. C'est cette implémentation que nous avons utiliserons en pratique. Elle a la particularité d'interrompre la progression dans le treillis des itemsets à ceux de taille 5.

En utilisant des seuils de support et de confiance respectivement de 30% et 60%, APRIORI génère 21 règles sur le jeu de données "jouet". Elles sont listées en table I.3.

id	prémisse	→	conclusion
1	g	→	a
2	a	→	g
3	g	→	d
4	b	→	f
5	f	→	d
6	b	→	e
7	e	→	b
8	b	→	d
9	a	→	c
10	a	→	e
11	a	→	d
12	c	→	e
13	e	→	c
14	c	→	d
15	e	→	d
16	g a	→	d
17	g d	→	a
18	a d	→	g
19	b e	→	d
20	b d	→	e
21	e d	→	b

TAB. I.3 – Règles extraites des données "jouet"

De nombreux travaux se sont attachés à l'amélioration de l'algorithme AIS, afin d'en améliorer les performances ou de généraliser son fonctionnement.

Nous venons de l'évoquer, dans [Agrawal and Srikant, 1994], les auteurs proposent l'algorithme APRIORI qui est celui sur lequel est basé l'outil d'extraction de règles que nous utiliserons. Dans cet article, ils proposent également deux autres algorithmes, APRIORITID et APRIORIHYBRID. Les deux premiers sont des améliorations d'AIS. D'une part, ils ne limitent plus la taille de la conclusion à un seul élément, et d'autre part, dans sa seconde version, APRIORITID n'effectue plus de passes sur les données afin de générer les itemsets candidats. Cette alternative est efficace afin d'extraire des itemsets de grande taille. Ces deux approches sont utilisées dans la proposition "hybride", combinant les deux stratégies.

Les algorithmes DHP (Direct Hashing and Pruning) [Park et al., 1995, Park et al., 1997] et IHP (Invert Hashing and Pruning) [Holt and Chung, 2002] focalisent sur l'étape de génération des candidats, et visent à réduire la taille des données à analyser lors de chaque passe. Dans [Brin et al., 1997b], les auteurs réordonnent les items de la base de cas afin de limiter les comptages nécessaires et proposent la méthode DIC (Dynamic Itemset Counting). OCD [Mannila et al., 1994], quant à lui, est fondé sur une analyse combinatoire de l'information issue des passes précédentes de l'algorithme, en vue d'écarter des candidats générés inutilement.

Des alternatives à AIS ont également été proposées. Ainsi, PARTITION [Savasere et al., 1995] permet de paralléliser le processus d'extraction de règles, et ainsi de traiter un volume de données plus important. Une autre approche parallélisée, MAXCLIQUE, est proposée par [Zaki et al., 1997]. Dans les algorithmes N-BM (Naïve BitMap) et H-BM (Hierarchical BitMap) proposés par [Gardarin et al., 1998], les auteurs développent une approche optimisant les coûts de calculs par la construction d'un index binaire sur les données.

Enfin, des généralisations du concept de règles d'association étendent les possibilités de traitement. Ainsi, [Guillaume, 2000] étend leur définition au cas des variables ordinales, et [Wu et al., 2004, Cornelis et al., 2006, Dong et al., 2006] proposent des méthodes permettant d'extraire des règles positives et négatives.

De par les nombreux travaux dans ce domaine, et l'amélioration des capacités de calcul, le volume de données traitées ne cesse d'augmenter. Ceci résulte en un nombre de règles extraites automatiquement impossible à valider manuellement. Une étape de sélection suite à l'extraction est alors nécessaire, afin d'orienter l'utilisateur dans cette base de connaissances.

Ainsi, l'une des étapes les plus importantes dans un processus d'ECD est la phase d'interprétation et d'évaluation des connaissances découvertes [Hilderman and Hamilton, 2000]. C'est particulièrement le cas dans les tâches d'apprentissage non supervisé, comme la découverte de règles d'association, où l'utilisateur ne spécifie généralement ni objectifs ni variables à expliquer.

Dans le cas des règles d'association, on doit faire face à deux problèmes : le premier est de nature quantitative – le nombre de règles est très élevé – et le second de nature qualitative – beaucoup de règles sont non satisfaisantes d'une point de vue statistique ou sans réel intérêt –.

Différentes solutions, non exclusives les unes des autres, ont été proposées pour traiter ces problèmes et aider l'utilisateur à explorer les règles *intéressantes*.

Une première approche consiste à réduire le nombre de règles. Cette réduction peut se réaliser en éliminant les règles redondantes [Lehn et al., 1998], [Zaki, 2000], [Zaman Ashrafi et al., 2004]. En permettant de filtrer les règles, les mesures d'intérêt complètent cette approche. C'est la voie que nous explorons dans ce mémoire.

Une seconde solution consiste à faire coopérer machine/algorithme et utilisateur en exploitant au mieux les capacités de ce dernier (capacités perceptives, cognitives). Pour tous les résultats numériques que nous avons proposé, nous avons dû aborder quelques problématiques de représentation visuelle afin de les restituer aisément et de façon synthétique. Des auteurs proposent ainsi des représentations graphiques des règles extraites [Lehn et al., 1999], [Pagès and Groetaers, 2003], [Hofmann and Wilhelm, 2001]. D'autres proposent des approches résolument ancrées dans la coopération homme/système [Kuntz et al., 2001], [Lehn, 2000], [Do and Poulet, 2003], [Blanchard, 2005].

Chapitre II

Mesures de qualité de règles d'association

Les algorithmes de type Ais vus au chapitre I fondés sur le support et la confiance ont tendance à produire un grand nombre de règles, tout spécialement lorsque les items sont fortement corrélés. Ce phénomène s'amplifie lorsque l'on diminue le seuil de support, ce qui est indispensable afin de découvrir des pépites. Le couple "support/confiance" ne permet pas de sélectionner toutes, et seulement toutes les règles intéressantes. Il est ainsi nécessaire de procéder à une post-analyse, cette dernière devant servir à retenir les règles pertinentes. Face au grand nombre de règles, il est impossible de prendre en compte chacune d'elles individuellement afin de la valider ou de la rejeter en tant que connaissance utile. Toutefois, nombreuses sont les règles que l'on peut prédire comme étant intéressantes (ou inintéressantes), étant donné un contexte applicatif particulier. Les mesures d'intérêt peuvent alors jouer un rôle essentiel ([Hilderman and Hamilton, 2001a, Tan et al., 2002]) en permettant de filtrer ou d'ordonner les règles extraites automatiquement et exhaustivement selon des critères "idéalement" adaptés aux besoins des utilisateurs [Lenca et al., 2003b]. Pour deux articles récents de synthèse sur la notion de qualité en Ecd, on peut par exemple se référer à [Hilderman and Hamilton, 2003, McGarry, 2005].

Ainsi, bien que non pensées pour évaluer les règles d'association, puisqu'antérieures aux algorithmes de la famille Ais, d'anciennes mesures trouvent un nouveau terrain d'application en Ecd. D'autres ont été introduites pour faire face à des besoins spécifiques. On peut donc recenser un très grand nombre de mesures de qualité dans la littérature (voir par exemple [Yao and Zhong, 1999, Guillet, 2004] pour une liste de plus de quarante mesures de qualité).

Nous définissons dans un premier temps le critère d'éligibilité des mesures que nous retenons dans un contexte d'évaluation de règles d'association. A partir de ce critère d'éligibilité, nous avons retenu un certain nombre de mesures que nous présentons dans ce chapitre. Nous nous restreignons aux mesures dites "objectives" [Freitas, 1999, Hilderman and Hamilton, 2000], par opposition aux mesures dites "subjectives" [Silberschatz and Tuzhilin, 1995, Liu et al., 2000]. Ces premières évaluent la qualité des règles à partir des grandeurs caractéristiques de chaque règle présentées au chapitre I. Elles peuvent donc être vues comme des fonctions de ces grandeurs. Pour chacune d'entre elles, nous évaluons quelques caractéris-

tiques, classiques dans le cadre d'étude de fonctions et intéressantes en ECD. Cette première approche intuitive sera formalisée et développée pour des propriétés propres à notre contexte au chapitre III.

Les mesures retenues sont issues de deux grandes familles de mesures objectives : d'une part, celles associées à une famille de règle à découvrir (les mesures descriptives), et d'autre part celles associées à une modélisation statistique (les mesures probabilistes). Nous illustrerons la suite de nos propos principalement à partir de vingt de ces mesures. La démarche suivie étant valable pour un autre ensemble de mesures. Par exemple, nous les comparerons selon des propriétés formelles au chapitre III, et selon une approche expérimentale au chapitre IV. Dans le chapitre III nous nous proposerons également de répondre selon deux approches à la problématique du choix d'une ou plusieurs mesures adaptées à l'utilisateur et au contexte.

1 Critère d'éligibilité des mesures de qualité

Nous nous sommes restreints dans ce travail aux mesures objectives, et strictement décroissantes en fonction de $n_{a\bar{b}}$. Deux raisons principales nous ont guidés. La première est orientée vers l'utilisateur, la seconde a des fondements méthodologiques.

Tout d'abord, nous souhaitons demander l'information la plus synthétique possible à l'utilisateur, par exemple, tolère-t-il les contre-exemples... Concrètement, il s'agira pour lui de spécifier sa problématique plutôt que d'exprimer des connaissances qu'il pourrait avoir *a priori*. Cela nous a conduits naturellement vers les mesures objectives. Ensuite, afin d'assurer une cohérence entre les règles mises en avant par une mesure et le recours à un algorithme de type AIS, nous avons retenu les mesures décroissantes avec $n_{a\bar{b}}$. En effet, ces algorithmes ne se basant que sur les co-occurrences, plus le nombre de contre-exemples $n_{a\bar{b}}$ à une règle est grand, moins elle doit être considérée d'"intérêt".

Les mesures objectives, contrairement aux mesures subjectives, ne nécessitent pas de connaissances *a priori* sur les données (coût de détermination d'un attribut, asymétrie du coût d'une erreur de prédiction...). Ces mesures étant des fonctions à valeurs réelles de n, n_a, n_b et n_{ab}, et de par les liens entre ces différentes grandeurs vues en section 2.1 du chapitre I, on s'autorisera, par souci de concision ou afin de mieux mettre en évidence le sens d'une mesure, à utiliser les autres grandeurs dans nos formules. En particulier, il est courant de représenter les courbes représentatives des mesures de qualité en fonction de $n_{a\bar{b}}$, le nombre de contre-exemples, ce que nous ferons en particulier en sections 3 et 4 de ce chapitre.

Dans [Silberschatz and Tuzhilin, 1995], un ensemble de mesures de qualité subjectives sont proposées. L'approche nécessite de disposer d'un ensemble de croyances fortes et faibles sur les données, et présuppose donc une relative maîtrise du domaine abordé, ce que nous ne souhaitons pas requérir. Pour quelques travaux menés sur les mesures subjectives, nous renvoyons le lecteur par exemple à [Liu et al., 1997, Padmanabhan and Tuzhilin, 1998]. Les mesures objectives sont, comme nous l'avons déjà signalé, basées sur des évaluations numériques des règles et ne nécessitent pas d'apport d'information de la part de l'utilisateur dans un tel traitement. Deux exemples classiques de mesures objectives sont le support et la confiance, présentées au chapitre I.

Des études, semblables à celle développée dans ce mémoire, ont été proposées. En particulier, [Tan et al., 2002] effectuent une comparaison expérimentale de vingt et une mesures, la plupart étant symétriques au sens de la négation de la conclusion (dans ce cas, $\mu(\mathtt{A} \rightarrow \mathtt{B}) = \mu(\mathtt{A} \rightarrow \bar{\mathtt{B}})$[1]), afin d'assister un utilisateur à choisir une mesure de qualité. Ces mesures symétriques ne nous semblent pourtant pas adaptées à la nature même de règles d'association, étant donné la l'asymétrie du problème mentionné précédemment (*i.e.* la non prise en compte de l'absence d'information). Les mesures proposées dans [Tan et al., 2002], ainsi que diverses mesures classiques, sont initialement décroissantes avec le nombre de contre-exemples, atteignent un point d'inflexion (généralement une valeur nulle, atteinte à l'indépendance ou à l'indétermination), puis deviennent croissantes, comme illustré en figure II.1.

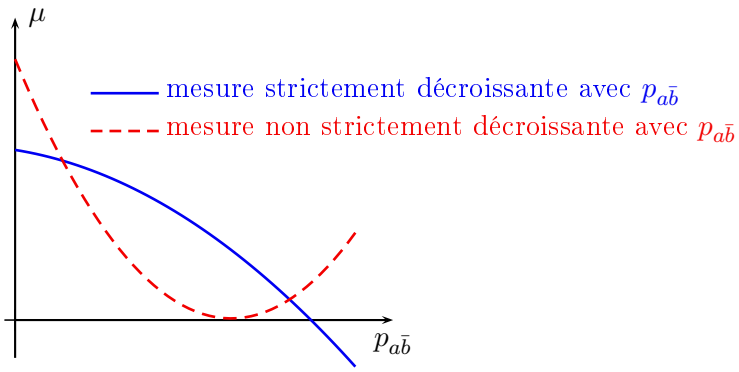


FIG. II.1 – Variations possibles d'une mesure de qualité en fonction de $p_{a\bar{b}}$

C'est ainsi que des mesures classiques, comme le coefficient du χ^2, la J-mesure [Goodman and Smyth, 1988], ou le coefficient de Gini [Gini, 1912], utilisées en induction d'arbres de décisions (paradigme d'apprentissage que nous n'aborderons pas dans ce mémoire ; le lecteur intéressé pourra se référer par exemple à [Breiman et al., 1984, Quinlan, 1992, Zighed and Rakotomalala, 2000]), ne seront pas utilisées dans notre étude. Notons de plus qu'à partir d'une mesure non symétrique, on peut se rapporter à une mesure symétrique, en prenant par exemple $\max\{\mu(\mathtt{A} \rightarrow \mathtt{B}), \mu(\mathtt{A} \rightarrow \bar{\mathtt{B}})\}$. Les mesures non symétriques au sens de la négation de la conclusion nous semblent ainsi plus informatives.

Afin de déterminer si une mesure est décroissante en fonction du nombre de contre-exemples, nous nous plaçons dans le cadre de la modélisation classique, et raisonnons à marges fixées (*i.e.* on considère la taille de l'antécédent et du conséquent comme étant constantes). Nous avons également proposé d'autres modélisations dans [Vaillant et al., 2006a]. Celles-ci sont présentées section 2.4 du chapitre III. La figure II.2 illustre cette modélisation.

Enfin, à l'exception d'une mesure probabiliste, toutes les mesures que nous retiendrons par la suite évaluent la qualité d'une règle de manière locale, *i.e.* sans prendre en compte les autres règles qui ont été ou auraient pu être générées par un algorithme.

[1] Bien que la règle $\mathtt{A} \rightarrow \bar{\mathtt{B}}$ ne fasse pas sens dans le cadre des règles d'association, nous notons par abus de notation $\mu(\mathtt{A} \rightarrow \bar{\mathtt{B}})$ la valeur que prend la mesure pour une règle ayant $n_{a\bar{b}}$ exemples, une prémisse de taille n_a, et une conclusion de taille $n_{\bar{b}}$.

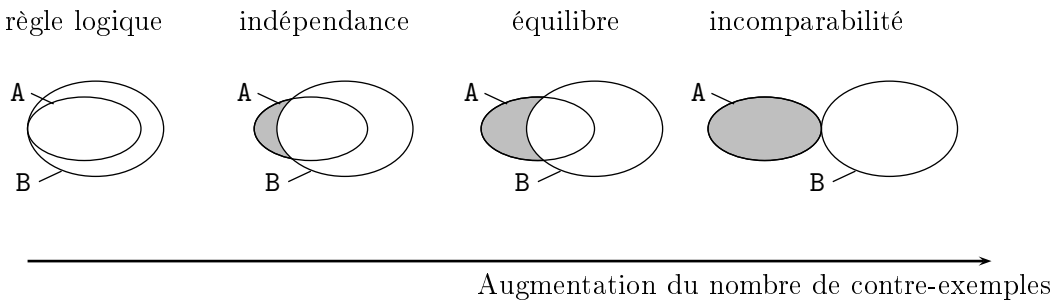


FIG. II.2 – Ajout de contre-exemples à une règle, modélisation classique

Parmi les caractéristiques classiques que l'on peut étudier, se trouvent les valeurs remarquables. Dans notre contexte, nous retiendrons la valeur maximale, et les valeurs prises à l'indépendance et à l'indétermination. L'allure de la courbe en fonction des contre-exemples pourra aussi nous intéresser. L'étude des valeurs prises par ces mesures en des références précises nous a conduits à proposer un cadre généralisé afin de quantifier la qualité de l'information extraite par des algorithmes d'extraction de règles d'association, que nous présentons dans un second temps.

2 Une "bonne" mesure de qualité : approche intuitive

Comme mentionné en section 1, nous avons choisi de nous restreindre aux mesures objectives, décroissantes avec les contre-exemples. On peut également proposer d'autres propriétés qui, lorsqu'elles sont respectées, rendent l'utilisation d'une mesure de qualité plus aisée.

En pratique, une mesure de qualité est souvent utilisée afin de filtrer un ensemble de règles, en plaçant un seuil d'acceptation minimal. C'est à l'utilisateur de choisir ce seuil. Selon les mesures, il peut être plus ou moins aisé de décider de sa valeur.

Dans le cas d'une mesure n'ayant pas une valeur maximale constante, on peut risquer de rejeter des règles potentiellement intéressantes (voir figure II.3 en fin de section).

En diminuant la valeur maximale de ce seuil afin d'éviter de rejeter les règles ayant peu de contre-exemples, on prend le risque d'accepter des règles ne faisant pas sens. Par exemple, des règles ne passant pas l'hypothèse d'indépendance (voir définition 5), ou des règles ne déviant pas suffisamment de l'équilibre (voir définition 6).

Définition 5 (Hypothèse d'indépendance) *On dit que deux itemsets* `A` *et* `B` *vérifient l'hypothèse d'indépendance (ou plus simplement qu'ils sont indépendants) lorsque les distributions des transactions vérifiant* `A` *et* `B` *sont statistiquement indépendantes.*

Cette hypothèse est vérifiée lorsque $p_{ab} = p_a \times p_b$.

Par exemple, sous hypothèse d'indépendance, la confiance d'une règle vaut p_b, et n'est donc pas constante. Ceci peut poser problème, tout particulièrement lorsque l'on sélectionne un ensemble de règles en fonction de CONF, en ne retenant que celles ayant une valeur

supérieure à un seuil donné. En effet, une règle à fort conséquent sera bien évaluée par CONF, même s'il n'y a pas de lien statistique dans les données. A l'opposé, rien ne garantit qu'en dessous d'une confiance donnée, les règles ne soient pas (statistiquement du moins) intéressantes.

Définition 6 (Ecart à l'équilibre [Blanchard et al., 2005a]) *On dit que deux itemsets* `A` *et* `B` *sont à l'équilibre lorsque la distribution des transactions vérifiant* `B` *est répartie de manière homogène sur celle des transactions vérifiant* `A`*. Il y a alors autant d'exemples que de contre-exemples à la règle* $\mathtt{A} \rightarrow \mathtt{B}$.

C'est le cas lorsque $p_{ab} = p_a/2$.

Par exemple, à l'équilibre, le support d'une règle vaut $p_a/2$, et n'est donc pas constant. Ceci peut à nouveau poser problème, lorsque l'on sélectionne un ensemble de règles en fonction de SUP, en ne retenant que celles ayant une valeur supérieure à un seuil donné. En effet, une règle à fort antécédent sera bien évaluée par SUP, même en deçà de l'équilibre.

Un autre désavantage à ne pas avoir de bornes fixes à l'indépendance et à l'équilibre est que les valeurs prises par une mesure de qualité pour ces situations de référence peuvent se croiser.

Nous illustrons cela, ainsi que la difficulté de placer un seuil en présence de valeurs de références non fixes au moyen de l'exemple suivant. Nous considérons deux familles de règles à marges fixées. Pour la première (que l'on notera $\mathtt{r_1}$), on fixe $p_a = 0.2$ et $p_b = 0.4$. Pour la seconde (que l'on notera $\mathtt{r_2}$), on fixe $p_a = 0.45$ et $p_b = 0.8$.

$\mathtt{r_1}$ a un support compris entre 0 et 0.2, l'indépendance se produit lorsque $p_{a\bar{b}} = 0.12$, et l'équilibre lorsque $p_{a\bar{b}} = 0.1$. $\mathtt{r_2}$ quant à elle a un support compris entre 0 et 0.45, l'indépendance se produit lorsque $p_{a\bar{b}} = 0.09$, et l'équilibre lorsque $p_{a\bar{b}} = 0.225$. Selon ces deux caractérisations, et en raisonnant à marges fixées, on voit déjà que l'ordre d'apparition des situations d'indépendance et d'équilibre en fonction des contre-exemples est inversé. La valeur prise par le support en ces valeurs de référence rend la fixation d'un seuil encore plus compliquée. Ces valeurs sont présentées dans le tableau II.1, et illustrées en figure II.3.

Si l'on utilise la famille de règles $\mathtt{r_1}$ pour fixer un seuil, que l'on considère la valeur prise par le support à l'indépendance ou à l'équilibre, on voit que les règles de la forme $\mathtt{r_2}$ dépassent largement ce seuil, même en deçà des situations qui, *a priori* ne nous intéressent plus. A l'inverse, si l'on considère la famille $\mathtt{r_2}$ afin de positionner un seuil, on élimine toute règle de type $\mathtt{r_1}$, même si elle n'a pas de contre-exemples ! La zone dénommée "conflit d'intérêt" dans la figure II.3 correspond aux valeurs de support *a priori* acceptables pour $\mathtt{r_1}$, et il conviendrait donc de placer un seuil d'acceptation en dessous de cette zone, mais également *a priori* non acceptables pour $\mathtt{r_2}$, et il conviendrait donc de placer un seuil d'acceptation au dessus de cette zone. Etant donné ce type de configuration, et pour SUP, quel que soit le choix de seuil, on sera soit trop sélectif, et on écartera des règles potentiellement intéressantes, soit pas assez, et on conservera des règles inintéressantes.

Enfin, connaître la valeur minimale prise par la mesure (dans ce cas, il n'y a pas d'exemples, les mesures étant décroissantes avec $n_{a\bar{b}}$) est un dernier repère qui permet de placer le seuil d'acceptation.

	$\mathtt{r}_1 : p_a = 0.2, p_b = 0.4$	$\mathtt{r}_2 : p_a = 0.45, p_b = 0.8$
règle logique	$p_{a\bar{b}} = 0$, SUP $= 0.2$	$p_{a\bar{b}} = 0$, SUP $= 0.45$
indépendance	$p_{a\bar{b}} = 0.12$, SUP $= 0.08$	$p_{a\bar{b}} = 0.09$, SUP $= 0.36$
équilibre	$p_{a\bar{b}} = 0.1$, SUP $= 0.1$	$p_{a\bar{b}} = 0.225$, SUP $= 0.225$

TAB. II.1 – Valeurs de SUP pour deux familles de règles

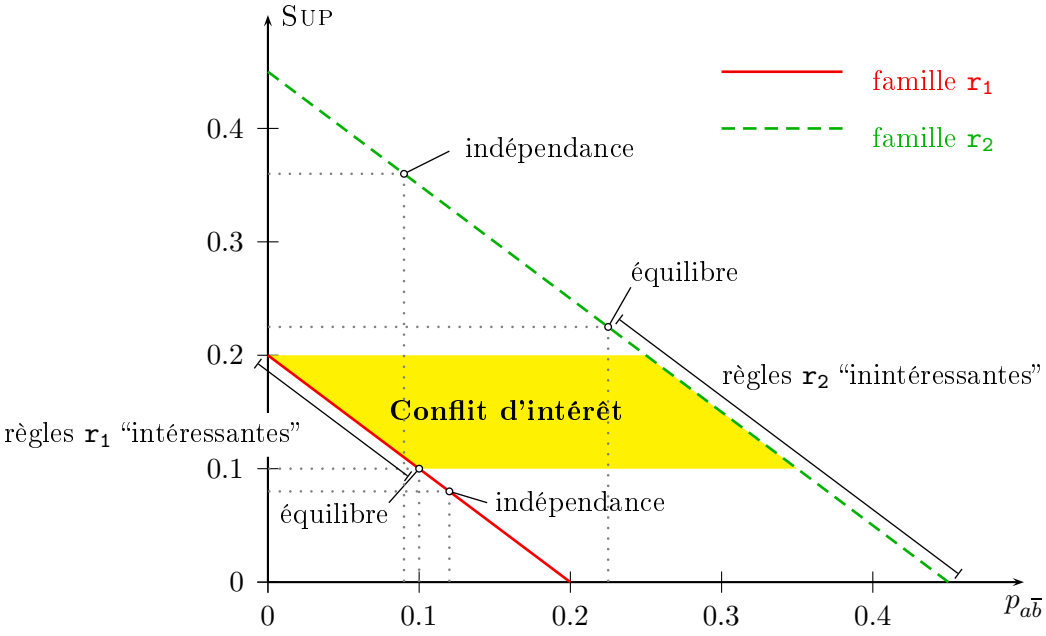


FIG. II.3 – Variations de SUP pour deux familles de règles

3 Mesures descriptives

Nous présentons dans cette section diverses mesures de qualité ayant été utilisées dans les études proposées dans ce mémoire. Nous donnons leur expression, ainsi que les valeurs qu'elles prennent pour les situations de références mentionnées section 2.

Afin d'illustrer nos propos, nous traçons les courbes représentatives de ces mesures en fonction de $n_{a\bar{b}}$ (ou $p_{a\bar{b}}$). Nous avons choisi trois grandeurs différentes pour p_a et p_b, pour $n = 2000$, que nous conserverons tout au long de ce chapitre :

- $p_a = 0.6$ et $p_b = 0.7$, l'indépendance se produit alors pour $p_{ab} = 0.42$, soit $p_{a\bar{b}} = 0.18$,
- $p_a = 0.2$ et $p_b = 0.6$, l'indépendance se produit alors pour $p_{ab} = 0.12$, soit $p_{a\bar{b}} = 0.08$,
- $p_a = 0.25$ et $p_b = 0.4$, l'indépendance se produit alors pour $p_{ab} = 0.1$, soit $p_{a\bar{b}} = 0.15$.

Dans ces trois cas de figure, $p_a < p_b$. Nous avons volontairement choisi ce cas de figure, puisqu'alors $\text{CONF}(\texttt{A} \rightarrow \texttt{B}) > \text{CONF}(\texttt{B} \rightarrow \texttt{A})$.

3.1 Support, confiance

Les deux mesures les plus classiques, et qui sont à la base des algorithmes de la famille AIS, sont le support, noté SUP et la confiance, notée CONF :

- $\text{SUP}(\texttt{A} \rightarrow \texttt{B}) = \frac{n_{ab}}{n} = p_{ab}$
- $\text{CONF}(\texttt{A} \rightarrow \texttt{B}) = \frac{n_{ab}}{n_a} = \frac{\text{SUP}(\texttt{A} \rightarrow \texttt{B})}{\text{SUP}(\texttt{A})} = \frac{p_{ab}}{p_a}$

Notons qu'initialement, le support était défini comme valant p_a. Cette définition est d'ailleurs celle proposée par défaut dans l'implémentation APRIORI [Borgelt and Kruse, 2002] que nous utilisons.

A l'indépendance, $\text{SUP} = p_a p_b$, $\text{CONF} = p_b$. A l'équilibre, $\text{SUP} = p_a/2$ et $\text{CONF} = 1/2$.

Le recours à ces deux mesures en tant que mesures de qualité de règles a toutefois été remis en question dans de très nombreux travaux (voir par exemple [Sese and Morishita, 2002, Lin and Tseng, 2006]). Toutefois, le support reste fortement utilisé pour des raisons algorithmiques. Il semble en effet difficile de se passer du support lors de la phase d'exploration du treillis d'itemsets, dont le rôle est de restreindre l'espace de recherche dans le treillis des itemsets (voir graphiques I.3 à I.6 du chapitre I pour quelques exemples de treillis). De par sa propriété d'antimonotonicité, évoquée précédemment, on peut optimiser le parcours de tels treillis. [Zimmermann and Raedt, 2004] utilisent la convexité de certaines mesures de qualité afin de procéder à une extraction de règles d'association en se passant de la condition de support minimal, mais cet apprentissage sur les données se fait de manière supervisée. D'autres approches semblent également prometteuses, pour autant aucune ne se démarque suffisamment pour que l'on puisse envisager de ne plus utiliser le support dans de nombreux algorithmes.

La confiance, quant à elle, se comprend intuitivement, ce qui constitue un atout majeur lorsque l'on présente un ensemble de règles à un expert et que leur qualité est quantifiée en fonction de CONF. En effet, correspondant à la probabilité conditionnelle d'avoir B sachant A, sa lecture est donc des plus directes.

3.2 Coefficient de corrélation

Le coefficient de corrélation de Pearson [Pearson, 1896], noté R, se définit comme :

$$\text{R} = \frac{p_{ab} - p_a p_b}{\sqrt{p_a p_b p_{\bar{a}} p_{\bar{b}}}}$$

Il permet de mesurer un écart à l'indépendance. Plus sa valeur est faible, plus `A` et `B` sont corrélés négativement. Plus sa valeur est grande, plus ils sont corrélés positivement. A l'indépendance, le coefficient de corrélation est nul. A l'équilibre, il vaut $\sqrt{\frac{p_a}{p_{\bar{a}} p_b p_{\bar{b}}}}(1/2 - p_b)$.

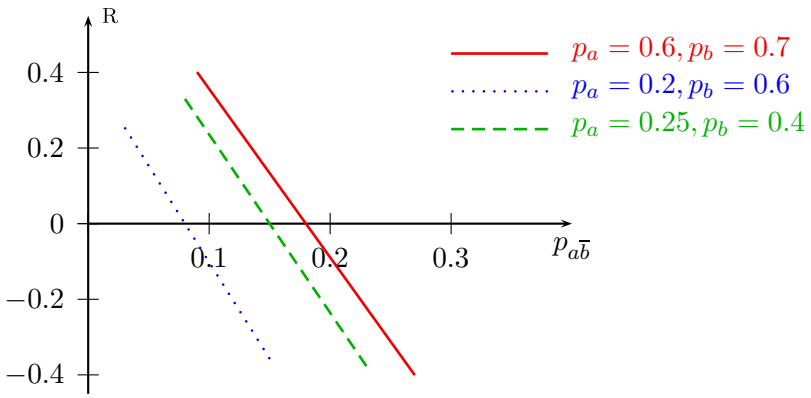


FIG. II.4 – Variations de R en fonction des contre-exemples

Les valeurs maximales de R dépendant des fréquences marginales, il n'est pas toujours simple de les interpréter.

3.3 Coefficient de Loevinger

Le coefficient H de Loevinger [Loevinger, 1947], que l'on note ici LOE, n'a pas ce défaut :

$$\text{LOE} = 1 - \frac{p_{a\bar{b}}}{p_a p_{\bar{b}}}$$

Sa valeur maximale est 1, lorsqu'il n'y a pas de contre-exemples à la règle, et il est nul à l'indépendance. A l'équilibre, il vaut $\frac{1/2 - p_b}{p_{\bar{b}}}$.

3.4 Confiance centrée

Diverses transformations de la confiance ont été proposées afin de pallier à son défaut de ne pas être constante à l'indépendance. Ceci est généralement fait en la recentrant par rapport à p_b.

Ainsi, la confiance centrée, CONFCEN, centre additivement la confiance autour de p_b :

$$\text{CONFCEN} = \text{CONF} - p_b = \frac{p_{ab} - p_a \times p_b}{p_a}$$

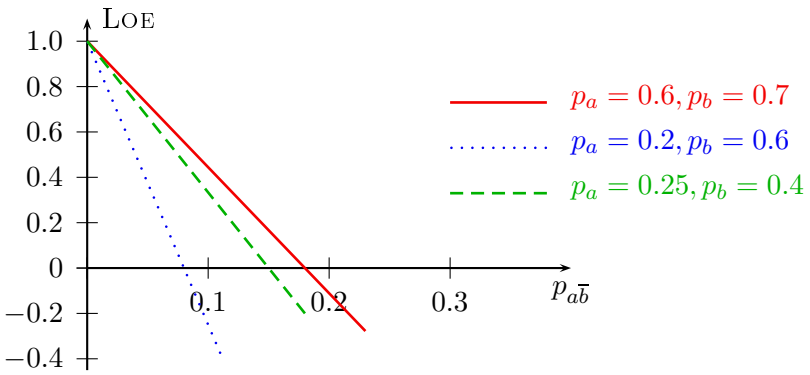


Fig. II.5 – Variations de Loe en fonction des contre-exemples

Sa valeur maximale est $p_{\bar{b}}$, et n'est donc plus constante, mais sa valeur est nulle sous hypothèse d'indépendance. A l'équilibre, elle vaut $1/2 - p_b$.

La figure II.6 présente les variations de ConfCen en fonction des contre-exemples.

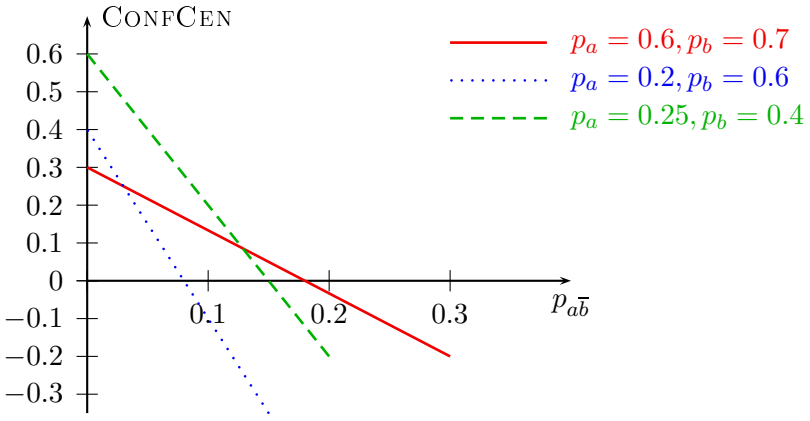


Fig. II.6 – Variations de ConfCen en fonction des contre-exemples

3.5 Conviction

Afin de quantifier l'implication, [Brin et al., 1997a] définissent la conviction comme suit :

$$\text{Conv} = \frac{p_a p_{\bar{b}}}{p_{a\bar{b}}}$$

Conv permet de quantifier l'écart à l'indépendance de $\mathtt{A} \wedge \bar{\mathtt{B}}$. Sa valeur est infinie dans le cas d'une implication logique, et est de 1 lorsque la prémisse et la conclusion de la règle sont statistiquement indépendants. A l'équilibre, elle vaut $2p_{\bar{b}}$.

Notons qu'à notre connaissance, c'est la première mesure ajoutée (avec le Lift) au couple Sup/Conf, afin de corriger les inconvénients de ce dernier, dans le cadre des règles d'association.

Ses variations sont représentées figure II.7. Remarquons que contrairement aux mesures précédemment présentées (linéaires selon $p_{a\bar{b}}$), sa décroissance est forte à l'apparition des premiers contre-exemples.

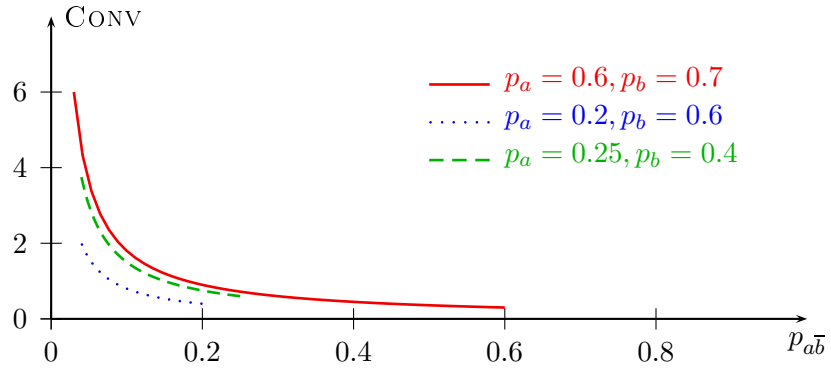


Fig. II.7 – Variations de Conv en fonction des contre-exemples

3.6 Estimateur de Laplace

Une mesure issue des méthodes Bayésiennes et utilisée en apprentissage, par exemple dans Cn2 [Clark and Niblett, 1989], est l'estimateur de Laplace [Good, 1965]. Elle estime la probabilité d'avoir B sachant A, et vaut sous sa forme la plus générale :

$$\text{Lap}(\mathtt{A} \rightarrow \mathtt{B}) = \frac{n_{ab} + 1}{n_a + k}$$

où k correspond au nombre de modalités que peut prendre B.

Dans le cas de données binaires, il n'y a que deux classes possibles ($k = 2$, et $\text{Lap}(\mathtt{A} \rightarrow \mathtt{B}) = \frac{n_{ab}+1}{n_a+2}$). Dès lors que n_{ab} (et donc n_a, puisque $n_a \geq n_{ab}$) est grand, Lap $\simeq$ Conf.

A l'indépendance, l'estimateur vaut $\frac{n_a p_b+1}{n_a+2}$, et 1/2 à l'équilibre.

3.7 Facteur Bayésien

FB [Jeffreys, 1935], quant à elle, est une forme d'odds ratio faisant intervenir les fréquences marginales de B :

$$\text{FB} = \frac{p_{ab} p_{\bar{b}}}{p_{a\bar{b}} p_b}$$

On peut ainsi l'interpréter comme le rapport entre la proportion de surface couverte par $\mathtt{A} \rightarrow \mathtt{B}$ dans B et la proportion de surface couverte par $\mathtt{A} \rightarrow \bar{\mathtt{B}}$ dans $\bar{\mathtt{B}}$ dans la figure I.2. Ce facteur vaut 1 à l'indépendance, et $p_{\bar{b}}/p_b$ à l'équilibre.

La figure II.8 en présente les variations.

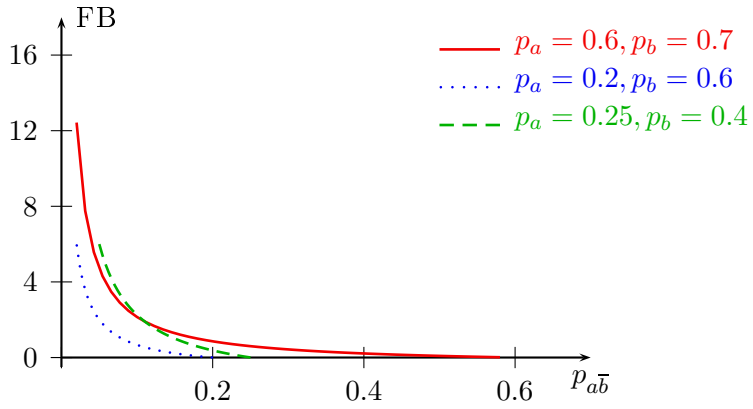


FIG. II.8 – Variations de FB en fonction des contre-exemples

3.8 Gain informationnel

Le gain informationnel (GI, [Church and Hanks, 1990], représenté figure II.9), est utilisé par exemple dans l'algorithme CART [Breiman et al., 1984] de construction d'arbres de décision. Issu de la théorie de l'information, GI correspond à la quantité d'information apportée par une règle et vaut :

$$\mathrm{GI} = \log(\frac{p_{ab}}{p_a p_b})$$

Il est nul à l'indépendance (il n'y a pas d'apport d'information), et vaut $-\log(2p_b)$ à l'équilibre.

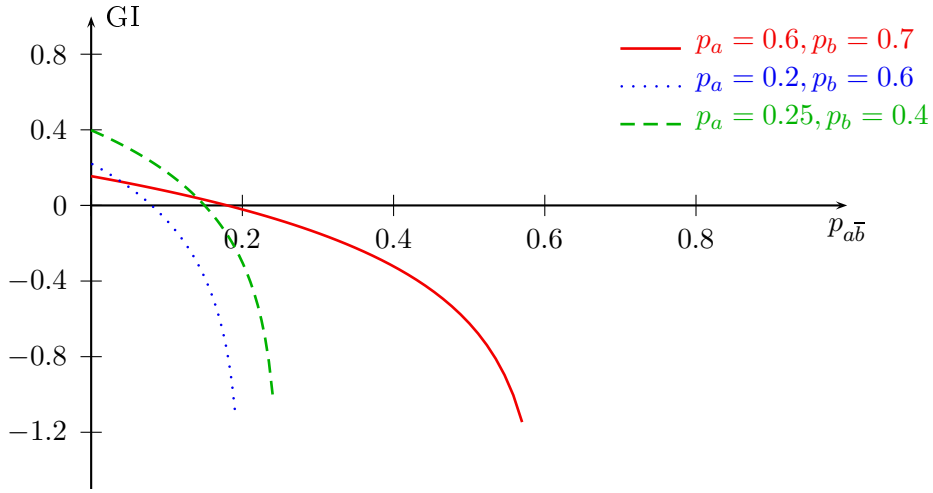


FIG. II.9 – Variations de GI en fonction des contre-exemples

3.9 Indice de qualité de Cohen

Une mesure issue de la théorie des tests est l'indice de qualité de Cohen ([Cohen, 1960]), ou Kappa, que nous notons ici IQC. Ce test est très fréquemment utilisé, en particulier en présence d'échelles binaires ou nominales, afin de de quantifier le niveau d'accord au sein d'un ensemble de votants, par delà le pur hasard. Il se définit comme suit, dans le cas d'un jugement à deux modalités :

$$\text{IQC} = 2\frac{p_{ab} - p_a p_b}{p_a p_{\bar{b}} + p_{\bar{a}} p_b}$$

Sa valeur est nulle à l'indépendance, et vaut $2\frac{p_a(1/2-p_b)}{p_a p_{\bar{b}}+p_{\bar{a}} p_b}$ à l'équilibre.

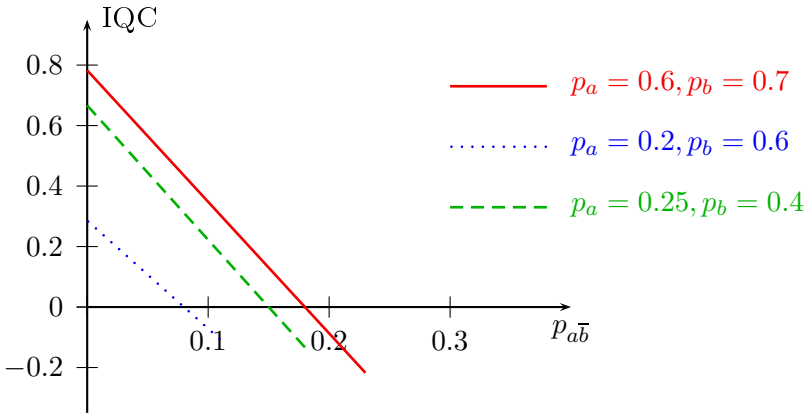


FIG. II.10 – Variations de IQC en fonction des contre-exemples

3.10 Lift

Si CONFCEN centre additivement la confiance, il est également possible de la centrer multiplicativement autour de p_b. Ce recentrage est ce qui est fait pour le lift [Brin et al., 1997a], noté LIFT :

$$\text{LIFT} = \text{CONF}/p_b = \frac{p_{ab}}{p_a p_b}$$

Sous hypothèse d'indépendance, LIFT a une valeur constante égale à 1. De même que dans le cas de CONFCEN, ce recentrage fait perdre la borne maximale constante, puisqu'en l'absence de contre-exemples, le LIFT vaut $1/p_b$. A l'équilibre, il vaut $\frac{1}{2p_b}$.

La figure II.11 présente ses variations en fonction du nombre de contre-exemples.

3.11 Mesure de Fukuda

La mesure de Fukuda [Fukuda et al., 1996] est une mesure prenant en compte le seuil de confiance σ_c utilisé lors de l'étape d'extraction de règles. Elle est définie comme :

$$\text{FUKUDA} = n_{ab} - \sigma_c n_a$$

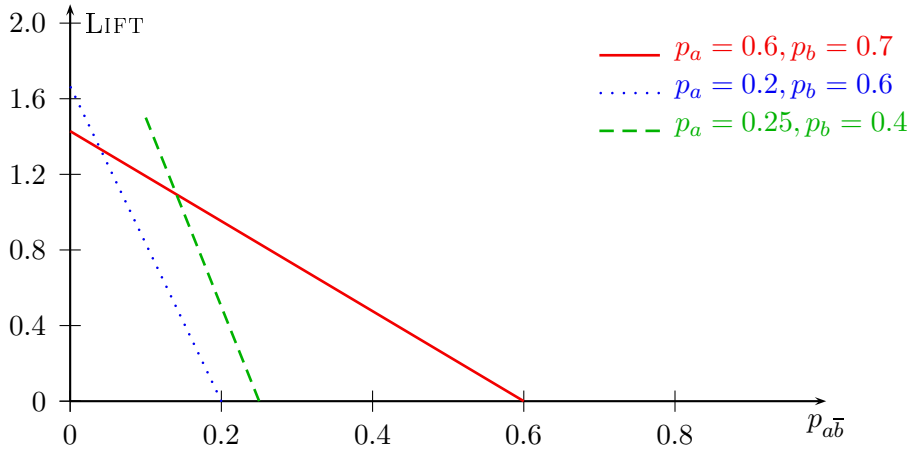


Fig. II.11 – Variations de Lift en fonction des contre-exemples

Elle prend donc ces valeurs dans l'intervalle $[0, n_a(1-\sigma_c)]$, vaut $n_a(p_b-\sigma_c)$ à l'indépendance, et $n_a(1/2-\sigma_c)$ à l'indétermination. Pour $\sigma_c = 0.6$, la figure II.12 présente les variations de Fukuda en fonction du nombre de contre-exemples.

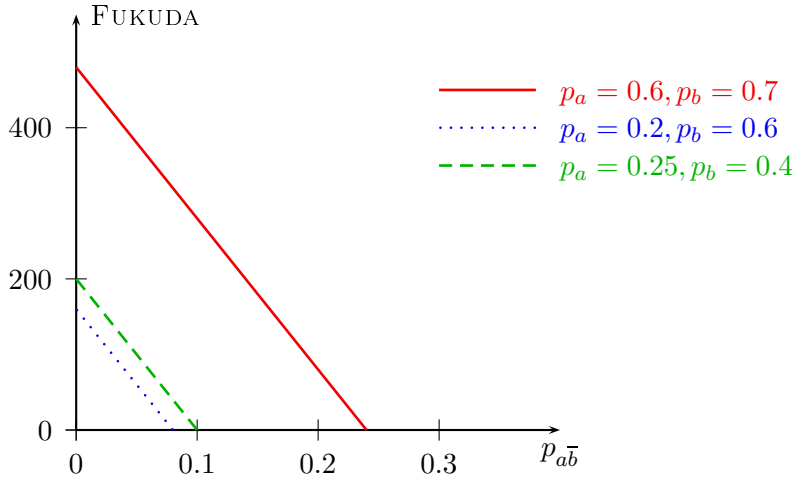


Fig. II.12 – Variations de Fukuda en fonction des contre-exemples ($\sigma_c = 0.6$)

3.12 Mesure de Ganascia

La mesure de Ganascia [Ganascia, 1987] est définie comme :

$$\text{Gan} = 2\text{Conf} - 1 = 2(p_{ab}/p_a - 0.5)$$

Sa valeur à l'indépendance est $2p_b - 1$, et elle est nulle à l'équilibre. Elle a été construite pour avoir des valeurs positives lorsqu'il y a plus d'exemples que de contre-exemples.

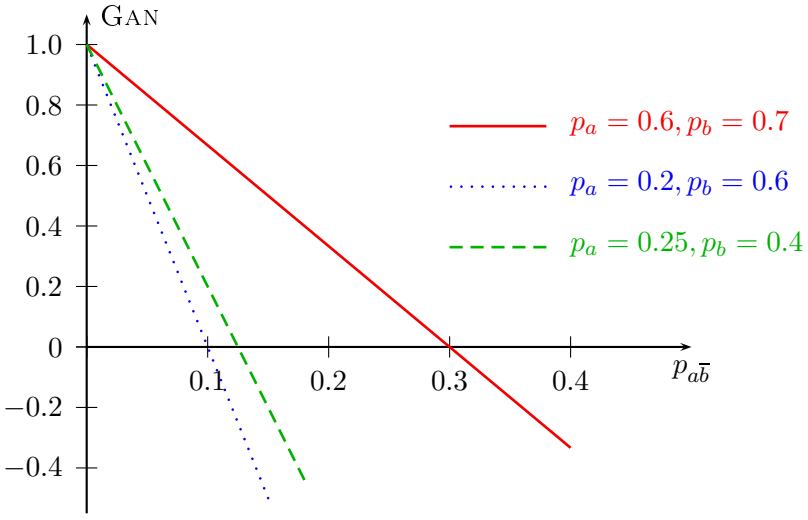


FIG. II.13 – Variations de GAN en fonction des contre-exemples

3.13 Mesure de Jaccard

La mesure de Jaccard [Jaccard, 1908] est définie comme :

$$\text{JAC} = \frac{p_{ab}}{p_{a\bar{b}} + p_b}$$

Sa valeur à l'indépendance est $\frac{p_a p_b}{p_a p_{\bar{b}} + p_b}$, et $\frac{p_a}{p_a + 2p_b}$ à l'équilibre.

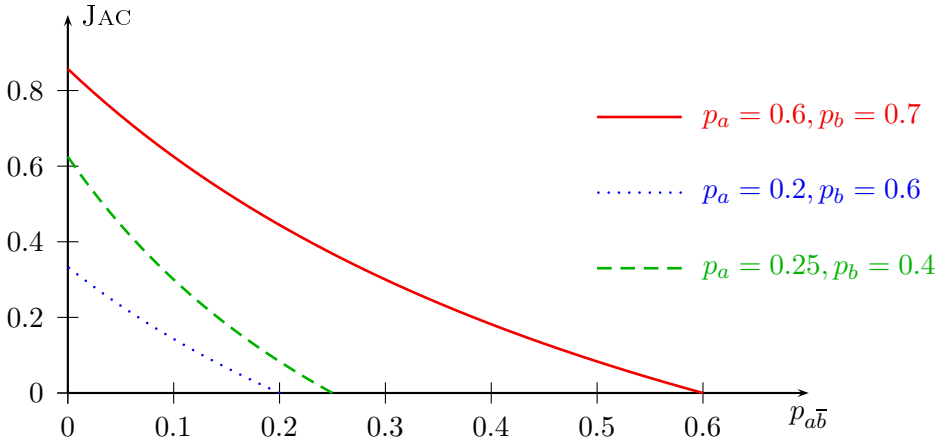


FIG. II.14 – Variations de JAC en fonction des contre-exemples

3.14 Mesure de Sebag-Schoenauer

La mesure proposée par [Sebag and Schoenauer, 1988] peut se lire comme une transformation de la confiance :

$$\text{SEB} = \frac{p_{ab}}{p_{a\bar{b}}} = \frac{1}{1/\text{CONF} - 1}$$

Bien que les deux mesures classent un ensemble de règles dans le même ordre, SEB se distingue de CONF de par son comportement (voir figure II.15). En effet, sa décroissance est forte à l'apparition des premiers contre-exemples. Ainsi, les règles n'ayant que peu de contre-exemples seront numériquement mises en avant.

A l'indépendance, SEB vaut $p_b/p_{\bar{b}}$. Elle vaut 1 à l'équilibre.

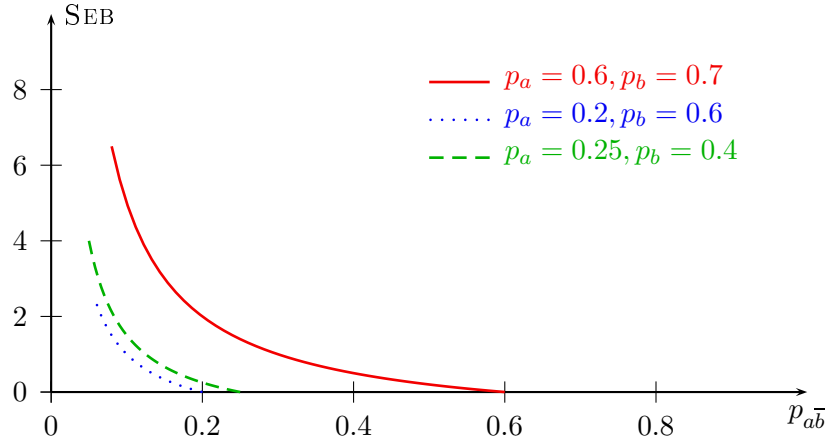


FIG. II.15 – Variations de SEB en fonction des contre-exemples

3.15 Mesure de Piatetsky-Shapiro

La mesure proposée par [Piatetsky-Shapiro, 1991] s'annule de même à l'indépendance :

$$\mathrm{PS}(\mathtt{A} \rightarrow \mathtt{B}) = n(p_{ab} - p_a p_b)$$

Elle a été construite en vue de posséder plusieurs des propriétés intéressantes qui seront discutées au chapitre III. Ses variations en fonction du nombre de contre-exemples sont représentées figure II.16.

Sa valeur est nulle à l'indépendance, et de $n_a(1/2 - p_b)$ à l'équilibre.

3.16 Moindre contradiction

La mesure proposée par [Azé and Kodratoff, 2002] est la moindre contradiction (MOCO, représentée figure II.17). L'objectif de cette mesure est de mettre en avant les règles ayant proportionnellement plus d'exemples que de contre-exemples, dans le cadre de la recherche de "pépites".

Celle-ci fait donc intervenir la différence entre exemples et contre-exemples et est normalisée selon la taille de la conclusion de la règle :

$$\mathrm{MOCO} = \frac{p_{ab} - p_{a\bar{b}}}{p_b}$$

A l'indépendance, elle vaut $p_a \frac{2p_b - 1}{p_b}$, et est nulle à l'équilibre.

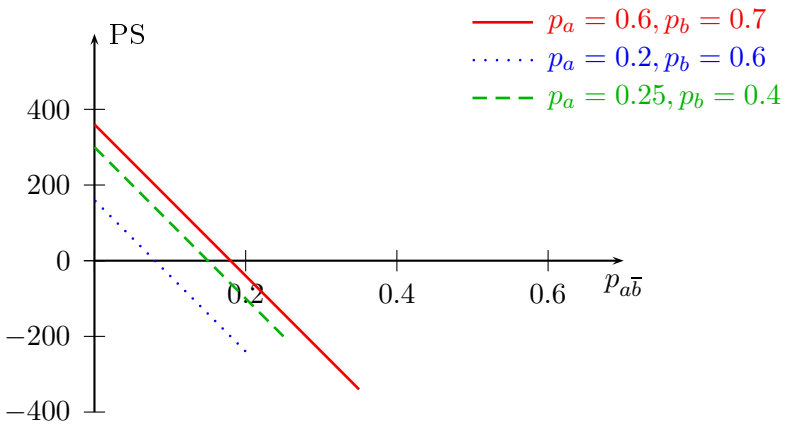


FIG. II.16 – Variations de PS en fonction des contre-exemples

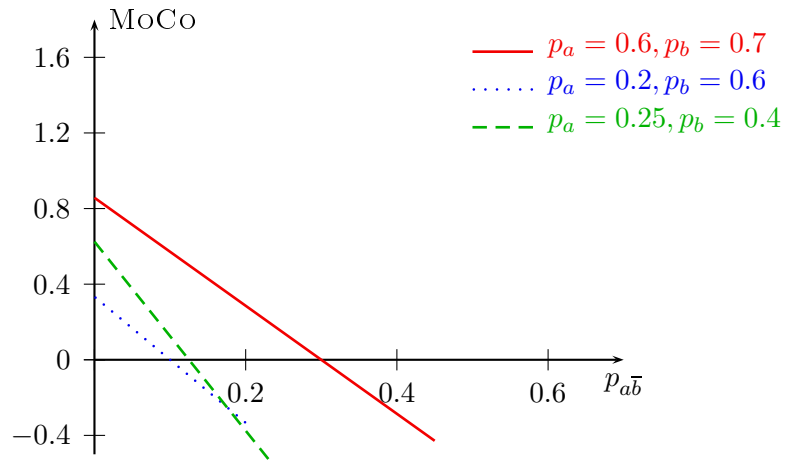


FIG. II.17 – Variations de MoCo en fonction des contre-exemples

3.17 Taux d'exemples et de contre-exemples

Classiquement utilisée en classification, on trouve le taux d'exemples et de contre-exemples, TEC :

$$\text{TEC} = \frac{p_{ab} - p_{a\bar{b}}}{p_{ab}}$$

Le taux vaut $\frac{2p_b - 1}{p_b}$ à l'indépendance, et est nul à l'équilibre.

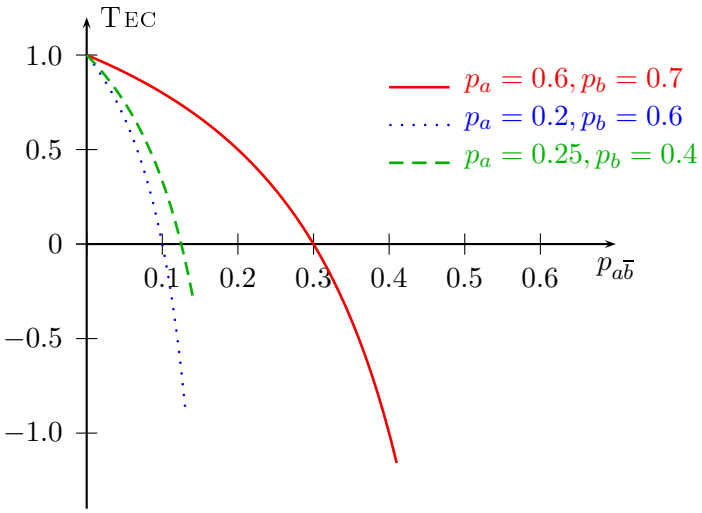


FIG. II.18 – Variations de TEC en fonction des contre-exemples

3.18 Zhang

Enfin, la mesure de ZHANG [Zhang, 2000], dont les variations sont présentées figure II.19, est définie comme suit :

$$\text{ZHANG} = \frac{p_{ab} - p_a p_b}{\max\{p_{ab}p_{\bar{b}}, p_b p_{a\bar{b}}\}}$$

Ce coefficient est nul à l'indépendance, et vaut $\frac{1-2p_b}{\max\{p_b, p_{\bar{b}}\}}$ à l'équilibre.

4 Indices statistiques et mesures probabilistes

A la base des mesures probabilistes, les indices statistiques proposés par [Lerman et al., 1981] correspondent à différentes modélisations visant à quantifier l'écart entre le nombre de contre-exemples observés par rapport à ce qui serait attendu, sous hypothèse d'indépendance. Trois niveaux distincts d'aléa sont proposés par les auteurs.

Reprenant leurs notations, on considère une règle $\mathtt{A} \rightarrow \mathtt{B}$, appliquée à un ensemble de n objets $\mathcal{O}$. $\mathcal{O}(A)$ (respectivement $\mathcal{O}(B)$) est le sous-ensemble d'objets de $\mathcal{O}$ vérifiant $\mathtt{A}$ (respectivement $\mathtt{B}$).

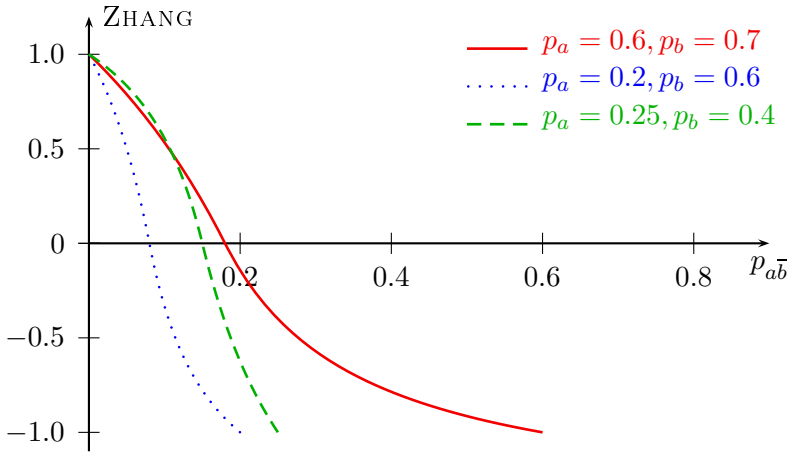


FIG. II.19 – Variations de ZHANG en fonction des contre-exemples

On se donne de plus deux ensembles aléatoires X et Y d'un ensemble Ω tels que l'espérance mathématique de la cardinalité de X (respectivement Y, Ω) soit n_a (respectivement n_b, n). On notera $\mathcal{N}$ le modèle (X, Y, Ω).

A partir des ensembles X et Y, on construit la variable aléatoire $\mathcal{S}$ comme étant $\mathcal{S} = card(X \cap Y)$. $\mathcal{S}$ est dénommée *indice brut aléatoire*. Etant donné cette variable aléatoire d'espérance $\mathcal{E}(\mathcal{S})$ et de variance $var(\mathcal{S})$, on construit l'indice q comme étant la normalisation de n_{ab} selon $\mathcal{S}$:

$$q(\mathtt{A}, \mathtt{B}) = \frac{n_{ab} - \mathcal{E}(\mathcal{S})}{\sqrt{var(\mathcal{S})}}$$

L'indice probabiliste local de vraisemblance du lien est alors défini comme suit :

$$\mathcal{I}(\mathtt{A}, \mathtt{B}) = P\{\mathcal{S} \leq n_{ab} | \mathcal{N}\} = P\{q(X, Y) \leq q(\mathtt{A}, \mathtt{B}) | \mathcal{N}\}$$

[Lerman et al., 1981] proposent alors trois formes d'hypothèse d'absence de liaison, notées $\mathcal{N}_1$, $\mathcal{N}_2$ et $\mathcal{N}_3$. Chaque modélisation met en relation un ensemble $\mathcal{L}$ de Ω avec un sous-ensemble $O \subseteq \mathcal{O}$, mais en utilisant différents niveaux d'aléa. $\mathcal{L}$ appartient donc à $\mathcal{P}(\Omega)$, l'ensemble des parties de Ω. On organise $\mathcal{P}(\Omega)$ en niveaux au sens de l'inclusion.

- Dans la modélisation $\mathcal{N}_1$, Ω est fixé, et sa cardinalité est n. $\mathcal{L}$ est choisi aléatoirement, sur le niveau associé à O.
- Dans la modélisation $\mathcal{N}_2$, Ω est toujours de cardinalité n, mais le niveau dans lequel on va effectuer le tirage de $\mathcal{L}$ est lui aussi choisi aléatoirement, selon la probabilité binomiale $C_n^k p(O)^k p(\bar{O})^{n-k}$, où $p(O) = |O|/n = n_o/n$ et $p(\bar{O}) = |\bar{O}|/n = n_{\bar{o}}/n$.
  A un niveau donné, le choix de $\mathcal{L}$ est effectué uniformément.
- Dans la modélisation $\mathcal{N}_3$, la cardinalité de Ω suit une loi de Poisson de paramètre n. On choisit ensuite le niveau et l'ensemble selon une méthode semblable à celle appliquée dans la modélisation $\mathcal{N}_2$.
  A Ω fixé, l'ensemble $\mathcal{L}$ n'est défini que pour $|\Omega| \geq |O|$. Dans ces conditions, posant $\gamma = |O|/|\Omega|$, le choix du niveau s'effectue avec la probabilité $C_n^k \gamma^k (1-\gamma)^{n-k}$.
  A un niveau donné, le choix de $\mathcal{L}$ est de même effectué uniformément.

[Lerman et al., 1981] démontrent dans ces conditions, que la distribution de $\mathcal{S}$ suit la loi :

- hypergéométrique (notée H) de paramètres n, n_a, p_b pour la modélisation $\mathcal{N}_1$,
- binomiale (notée B) de paramètres n, $p_a.p_b$ pour la modélisation $\mathcal{N}_2$,
- de Poisson (notée Poi) de paramètre $n.p_a.p_b$ pour la modélisation $\mathcal{N}_3$.

La table II.2 résume ces résultats, où q_i est l'indice q associé à la modélisation $\mathcal{N}_i$.

	$\mathcal{N}_1$	$\mathcal{N}_2$	$\mathcal{N}_3$
Loi de $\mathcal{S}$	$H(n, n_a, p_b)$	$B(n, p_a p_b)$	$Poi(np_a p_b)$
$\mathcal{E}(\mathcal{S})$	$n_a p_b$	$np_a p_b$	$np_a p_b$
$var(\mathcal{S})$	$\frac{n_{\bar{a}}}{n-1} n_a p_b p_{\bar{b}}$	$np_a p_b(1 - p_a p_b)$	$np_a p_b$
$q(\mathtt{A},\mathtt{B})$	$q_1(\mathtt{A},\mathtt{B}) = \sqrt{n-1}\frac{p_{ab}-p_a p_b}{\sqrt{p_{\bar{a}} p_a p_b p_{\bar{b}}}}$	$q_2(\mathtt{A},\mathtt{B}) = \sqrt{n}\frac{p_{ab}-p_a p_b}{\sqrt{p_a p_b(1-p_a p_b)}}$	$q_3(\mathtt{A},\mathtt{B}) = \sqrt{n}\frac{p_{ab}-p_a p_b}{\sqrt{p_a p_{\bar{b}}}}$

TAB. II.2 – Résumé des différentes modélisations

4.1 Indice d'implication

Selon $\mathcal{N}_3$, la modélisation des contre-exemples suit donc une loi de poisson de paramètre $n_a\ p_{\bar{b}}$, formant ainsi l'indice d'implication [Gras, 1979, Lerman et al., 1981] :

$$q_3(\mathtt{A},\bar{\mathtt{B}}) = \textsc{IndImp} = \sqrt{n}\frac{p_{a\bar{b}} - p_a p_{\bar{b}}}{\sqrt{p_a p_{\bar{b}}}}$$

Il faut donc considérer son opposé, -IndImp, si l'on souhaite évaluer la qualité d'une règle puisque IndImp est d'autant plus grand que le nombre de contre-exemples est important par rapport à une situation d'indépendance. L'indice est nul à l'indépendance, et vaut $\sqrt{\frac{np_a}{p_{\bar{b}}}}(1/2 - p_b)$ à l'indétermination.

La figure II.20 présente les variations de -IndImp en fonction de $p_{a\bar{b}}$.

4.2 Intensité d'implication

A partir de la modélisation $\mathcal{N}_3$, [Gras et al., 1996] construisent une mesure probabiliste, l'intensité d'implication (IntImp). Cette modélisation est préférée à la première car l'indice q_3 en dérivant est non-symétrique (*a contrario*, $q_1(\mathtt{A},\mathtt{B}) = q_1(\mathtt{B},\mathtt{A})$). D'autres travaux [Gras, 1979] indiquent que les résultats obtenus en considérant les modélisations $\mathcal{N}_2$ et $\mathcal{N}_3$ sont assez proches.

A partir de la définition de l'indice d'implication, on peut construire une mesure probabiliste, comme étant l'étonnement d'observer une valeur de l'indice statistique allant autant dans le sens de l'intérêt de la règle :

$$P\Big[poisson\big(np_a p_{\bar{b}}\big) \geq p_{a\bar{b}}\Big]$$

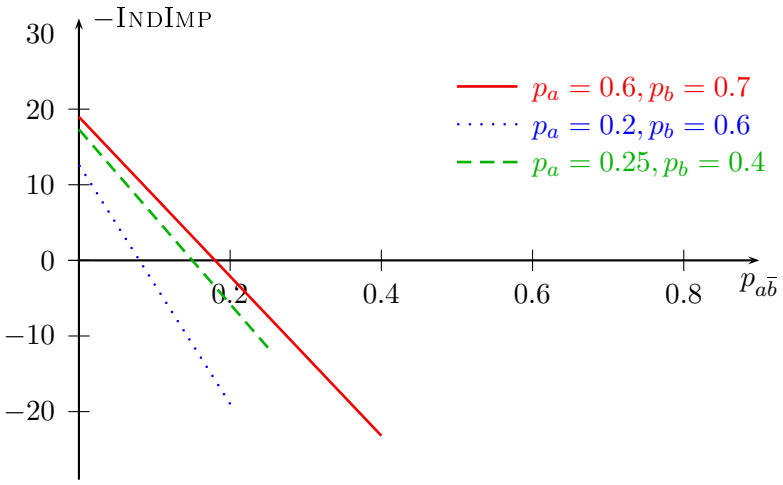


Fig. II.20 – Variations de -IndImp en fonction des contre-exemples

Sous des conditions légitimant l'approximation (paramètre de la loi supérieur à 10), il est possible d'approcher cette loi par la loi normale centrée réduite, et donc d'écrire, en effectuant la même opération sur $p_{a\bar{b}}$:

$$\text{IntImp} = P\Big[N(0,1) \geq \text{IndImp}\Big] = 1 - \Phi(\text{IndImp})$$

où Φ est la fonction de répartition de la loi normale centrée réduite.

Cette mesure vaut alors 1/2 à l'indépendance, et $P\Big[N(0,1) \geq \sqrt{\frac{np_a}{p_{\bar{b}}}}(1/2 - p_b)\Big]$ à l'équilibre, et vaut 1/2 pour $p_b = 1/2$.

La figure II.21 présente les variations de IntImp en fonction de $p_{a\bar{b}}$. Notons que cette mesure possède une propriété très intéressante puisqu'elle permet de tolérer l'apparition d'un certain nombre de contre-exemples avant de dégrader l'évaluation d'une règle.

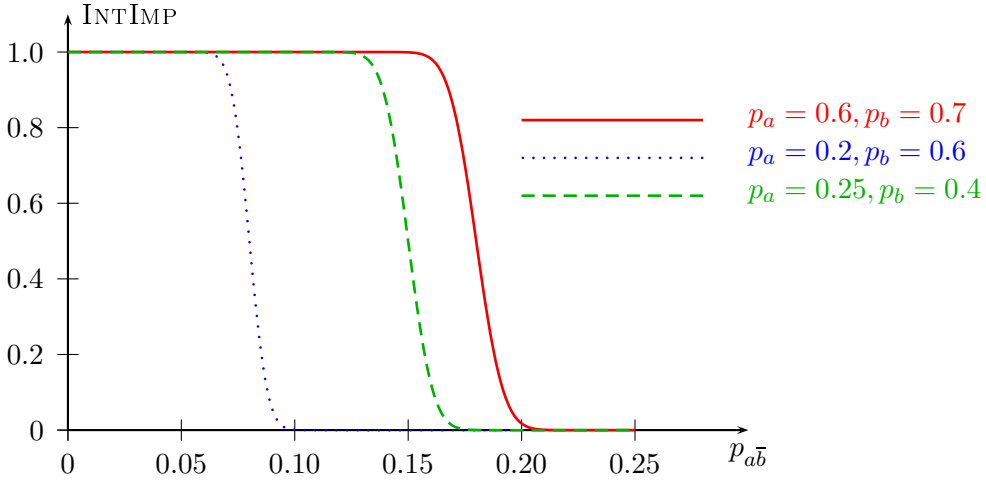


Fig. II.21 – Variations de IntImp en fonction des contre-exemples

4.3 Intensité d'implication entropique

Les auteurs reprochent toutefois à IntImp de rapidement tendre vers 1 ou 0 lorsque n devient grand (voir figure II.22). Ils proposent ainsi dans [Gras et al., 2001] de moduler IntImp au moyen d'un coefficient entropique, afin qu'elle conserve un pouvoir discriminant, même lorsque n croît. C'est ce coefficient entropique (non dépendant de n) qui corrige IntImp lorsque celle-ci n'est plus discriminante et que toutes les règles sont évaluées à 1. L'intensité d'implication entropique (représentée figure II.23) est donc définie comme :

$$\left\{[(1-h_1(p_{a\bar{b}})^2)\times(1-h_2(p_{a\bar{b}})^2)]^{1/4}\text{IntImp}\right\}^{1/2}$$

avec :

$$h_1(t) = -(1-\frac{t}{p_a})\log_2(1-\frac{t}{p_a}) - \frac{t}{p_a}\log_2(\frac{t}{p_a}) \quad \text{pour } t \in [0, p_a/2[$$
$$h_1(t) = 1 \quad \text{sinon}$$
$$h_2(t) = -(1-\frac{t}{p_{\bar{b}}})\log_2(1-\frac{t}{p_{\bar{b}}}) - \frac{t}{p_{\bar{b}}}\log_2(\frac{t}{p_{\bar{b}}}) \quad \text{pour } t \in [0, p_{\bar{b}}/2[$$
$$h_2(t) = 1 \quad \text{sinon}$$

On peut observer l'effet de l'introduction de ce coefficient en comparant les figures II.21 et II.23, où l'on distingue clairement dans la deuxième l'effet annulateur du coefficient entropique, dès lors que le nombre de contre-exemples devient trop important. De plus, en reprenant la figure II.22 et en traçant IIE, on observerait des seuils maximaux atteints différents pour chacune des trois courbes : le coefficient entropique a donc un effet de seuillage sur la qualité maximale que peuvent avoir les règles.

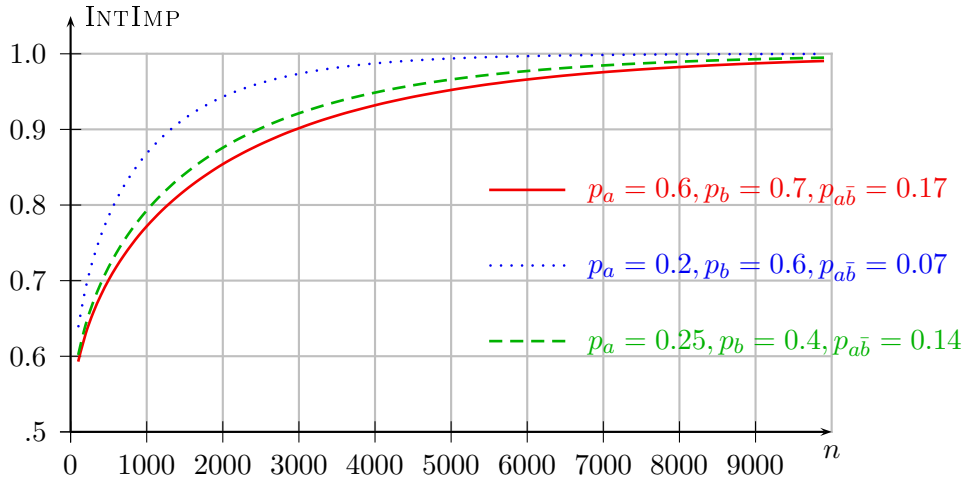


Fig. II.22 – Variations de IntImp en fonction de n

Bien que IntImp ait une valeur fixe à l'indépendance, IIE n'en a pas. Ceci est dû au coefficient entropique, qui ne prend pas de valeur constante à l'indépendance. Il est bien nul pour $p_{a\bar{b}} \geq 1/4$, mais ne l'est pas pour $p_{a\bar{b}} < 1/4$ lorsque $p_a < 1/2$ et $p_b > 1/2$.

Par contre, sa valeur est nulle à l'équilibre, puisqu'alors $h_1(t) = h_2(t) = 1$ par définition.

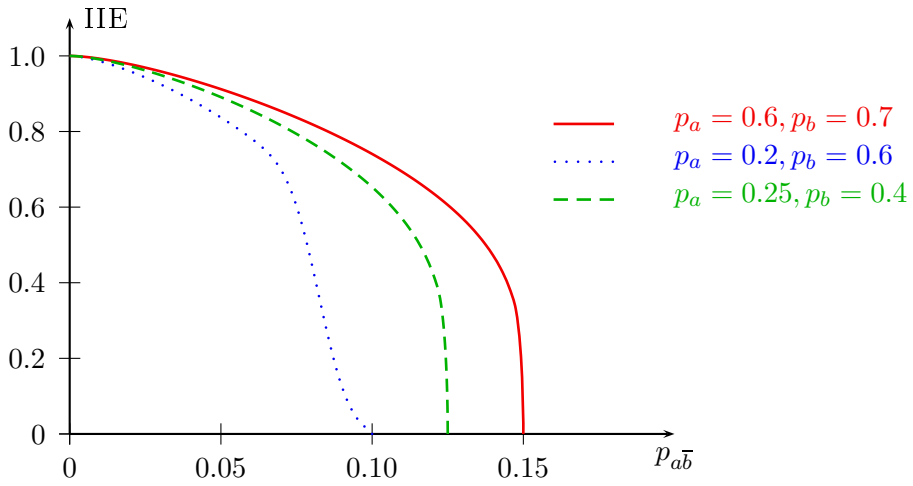


FIG. II.23 – Variations de IIE en fonction des contre-exemples

4.4 Intensité d'implication entropique tronquée et révisée

Afin qu'IIE ait une valeur fixe (nulle) à l'indépendance, nous avons proposé dans [Lallich et al., 2005b] deux modifications de sa définition. Dans les deux cas, il s'agit d'annuler l'un des coefficients intervenant dans sa définition. Une première approche consiste à recentrer l'intensité d'implication par rapport à 1/2, et l'autre à modifier la définition du coefficient entropique.

Annuler la partie faisant intervenir INTIMP à l'indépendance peut se faire par une transformation affine, que l'on borne inférieurement aux valeurs positives. Il s'agit donc d'une version révisée, et nous remplaçons INTIMP par $\max\{2\text{INTIMP} - 1; 0\}$ dans la formule de IIE, et appelons cette nouvelle mesure IIE révisée (RIIE). Le désavantage que nous voyons à cette méthode est que le spectre de valeurs d'IIE est alors modifié, comme on peut le voir sur les figures II.24 et II.25. La première figure compare ces deux mesures pour $n = 2000$, et sauf dans le cas où $p_a = 0.2$ et $p_b = 0.6$, la courbure de IIE est uniquement liée au travail du coefficient entropique, d'où des courbes quasiment sur la première bissectrice. Les points notés en rouge et vert correspondent à la dernière valeur non nulle d'IIE. En bleu, en partant du haut à droite vers le bas à gauche, sont notés :

- le début de la différence visible entre IIE et RIIE,
- le dernier point non nul de RIIE,
- le premier point nul de RIIE,
- le premier point où IIE et RIIE sont toutes les deux nulles.

En réduisant n à 200 (figure II.25), on augmente les variations de INTIMP, et ainsi on peut observer une légère incurvation des courbes rouges et vertes, mais ces différences restent faibles. Les différences majeures apparaissent dans le cas que l'on souhaitait corriger, à savoir $p_a < 1/2$ et $p_b > 1/2$.

La seconde approche consiste à changer les bornes annulant le coefficient entropique. Les conditions portant sur $p_{a\bar{b}}$ dans h_1 et h_2 que nous proposons sont alors respectivement $p_{a\bar{b}} < \min\{\frac{p_a}{2}; p_a p_{\bar{b}}\}$ et $p_{a\bar{b}} < \min\{\frac{p_{\bar{b}}}{2}; p_a p_{\bar{b}}\}$.

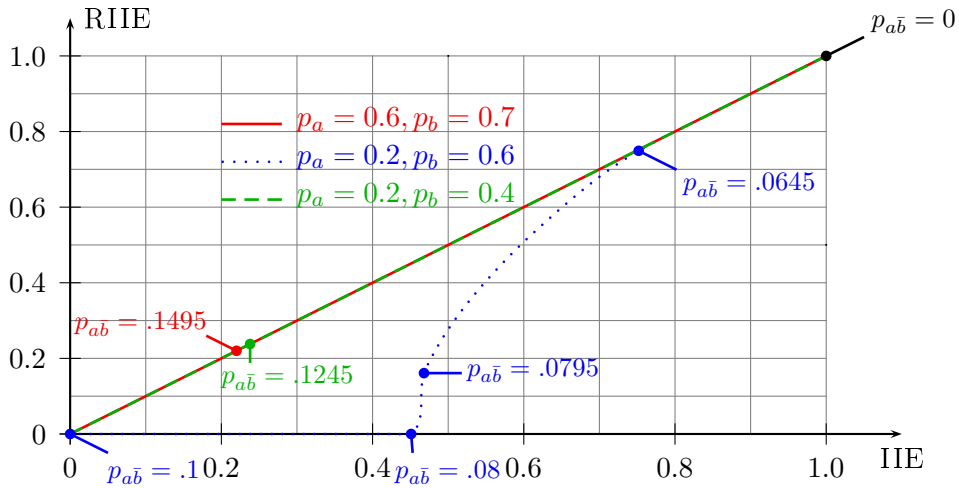


FIG. II.24 – Variations croisées de IIE et RIIE, $n = 2000$

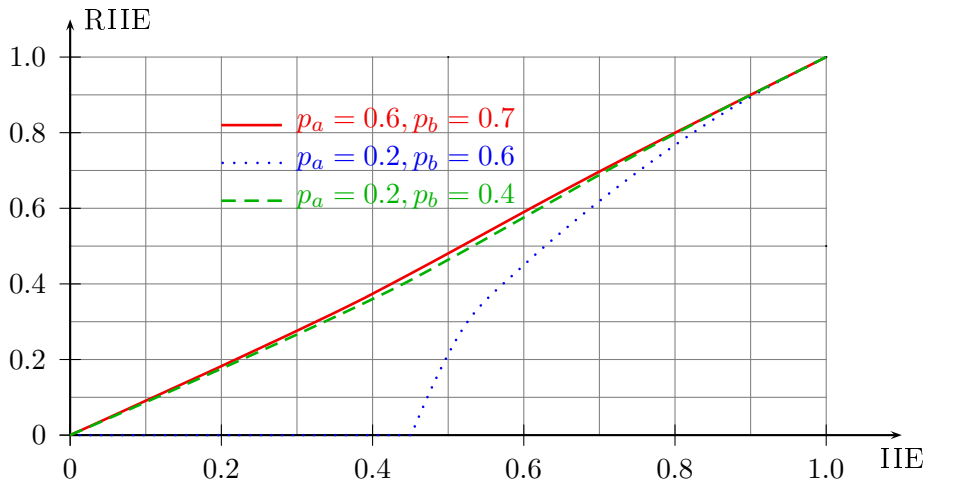


FIG. II.25 – Variations croisées de IIE et RIIE, $n = 200$

S'il n'est peut-être pas évident à première vue de faire le lien entre ces nouvelles bornes et les anciennes, une réécriture des conditions permet de mieux voir la logique (quelque peu masquée par des jeux d'écritures) suivie pour la détermination de ces nouvelles bornes.

La première condition de h_1, à savoir $p_{a\bar{b}} < \frac{p_a}{2}$, s'écrit également $p_{b|a} > 1/2$. Sous hypothèse d'indépendance, $p_{b|a} = p_b$. Si l'on souhaite imposer une valeur de 1 à h_1 lorsque l'on est en dessous de l'hypothèse d'indépendance, il suffit de modifier la dernière condition portant sur $p_{b|a}$ en $p_{b|a} > \max\{1/2; p_b\}$. Ainsi, dès lors que l'on est en dessous de l'hypothèse d'indépendance, h_1 vaut 1, ce qui a pour effet d'annuler le coefficient entropique. Si l'on revient aux notations en fonction de $p_{a\bar{b}}$, on retrouve bien la limite proposée. De même, la borne proposée pour h_2 vient d'une réécriture de la condition $p_{\bar{a}|\bar{b}} > 1/2$ (équivalente à $p_{a\bar{b}} < \frac{p_{\bar{b}}}{2}$) en $p_{\bar{a}|\bar{b}} > \max\{1/2; p_{\bar{a}}\}$.

Clairement, modifier une seule des conditions est suffisante à annuler le coefficient entropique sous hypothèse d'indépendance, mais nous préférons introduire ces deux modifications par souci de cohérence.

Nous appelons cette nouvelle intensité d'implication entropique l'intensité d'implication entropique tronquée, et la notons TIIE.

Elle est nulle à l'indépendance et à l'équilibre [Lallich et al., 2005a].

4.5 Indice probabiliste discriminant

Suivant la logique probabiliste que nous venons de décrire, [Lerman and Azé, 2003] proposent une mesure contextuelle. L'indice probabiliste discriminant, IPD, est également une référence à l'écart du nombre de contre-exemples basée sur IndImp. Toutefois, afin de prendre en compte la typologie des autres règles, IndImp est centré-réduit en fonction des valeurs prises par IndImp sur l'ensemble de la base de règles $\mathcal{B}$ évaluée (notation $^{CR/\mathcal{B}}$). Ainsi, on a :

$$\text{IPD} = P\Big[\mathcal{N}(0,1) > \text{IndImp}^{CR/\mathcal{B}}\Big]$$

C'est la seule mesure, parmi celles que nous avons étudiées, qui prenne en compte l'évaluation des autres règles de la base extraite. Ses valeurs pour les situations de référence dépendent alors des grandeurs utiles au centrage et à la réduction de IndImp. Notant m_{ii} la valeur moyenne d'IndImp, et e_{ii} son écart-type, l'indice vaut alors :

- $P\Big[N(0,1) \geq -\frac{m_{ii}}{e_{ii}}\Big]$ à l'indépendance,
- $P\Big[N(0,1) \geq \frac{\sqrt{\frac{np_a}{p_{\bar{b}}}}(1/2-p_b)-m_{ii}}{e_{ii}}\Big]$ à l'indétermination.

Ainsi, pour une base de règle donnée, la valeur d'IPD est fixe à l'indépendance, mais variable selon les bases de règles évaluées.

4.6 Indice probabiliste d'écart à l'équilibre

Dans [Blanchard et al., 2005a], suivant à nouveau une logique similaire à celle permettant de construire les mesures statistiques précédentes, les auteurs proposent un indice probabiliste d'écart à l'équilibre, IPEE. Cet indice se base sur la modélisation du nombre

de contre-exemples à une règle, sous l'hypothèse H_0 d'équilibre, *i.e.* $n_{ab} = n_{a\bar{b}}$. Sous cette condition, la variable aléatoire $\mathcal{N}_{a\bar{b}}$ suit une loi binomiale de paramètres n_a et 1/2. Pour $n_a/2$ supérieur à 5, on peut approcher la loi binomiale par la loi normale de moyenne $n_a/2$ et d'écart-type $\sqrt{n_a}/2$, soit :

$$\text{IPEE} = P\Big[N(0,1) \geq \frac{n_{a\bar{b}} - n_{ab}}{\sqrt{n_a}}\Big]$$

Sous hypothèse d'indépendance, IPEE vaut alors $P\Big[N(0,1) \geq \sqrt{n_a}(1-2p_b)\Big]$. Pour une règle logique, la mesure vaut $\sqrt{1-\frac{1}{2^{n_a}}}$, a une valeur constante égale à 1/2 à l'équilibre, et est nulle lorsqu'il n'y a pas d'exemple.

La figure II.26 présente les variations d'IPEE en fonction des contre-exemples. La figure II.27 illustre le fait qu'IPEE vaut 1/2 à l'équilibre (on s'intéresse ici à ses variations en fonction de la confiance, $p_{b|a}$, en faisant varier un paramètre γ, tel que $n_a = 20 \times \gamma$, $\gamma \in \{1; 5; 40; 1000\}$).

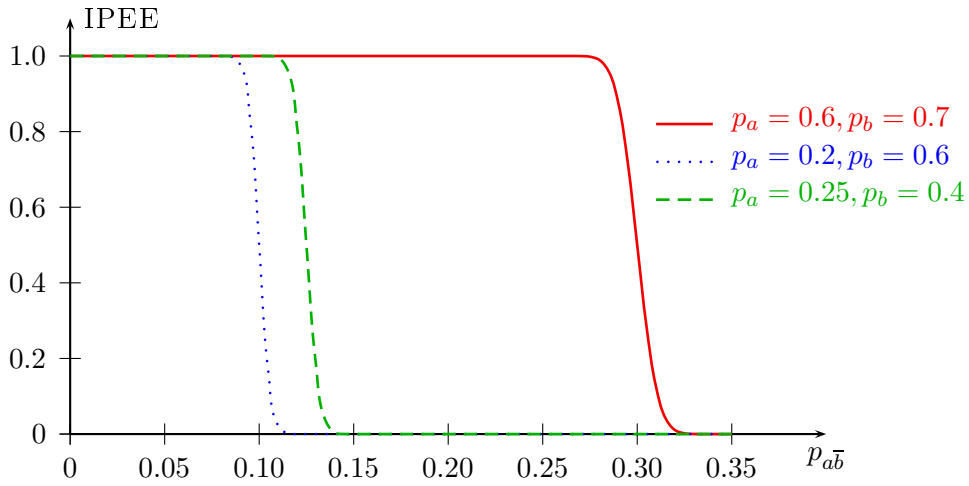


FIG. II.26 – Variations de IPEE en fonction des contre-exemples

4.7 Indice probabiliste d'écart à l'équilibre entropique

IPEE étant construite de manière semblable à INTIMP, elle souffre du même défaut, et devient peu discriminante lorsque n est grand. Dans [Blanchard et al., 2005b], les auteurs proposent donc de pondérer IPEE par l'indice d'inclusion, construisant ainsi la version entropique de cette mesure, IP3E. Toutefois, afin de moins dénaturer la valeur prise par IPEE dans sa version entropique, un recentrage sur 1/2 est effectué :

$$\text{IP3E} = \sqrt{\frac{1}{2}\big\{[(1-h_1(p_{a\bar{b}})^2) \times (1-h_2(p_{a\bar{b}})^2)]^{1/4} + 1\big\} \times \text{IPEE}}$$

Le recentrage de l'indice d'inclusion sur 0.5 a pour but de conserver une même valeur constante à l'équilibre pour IPEE et IP3E, $\frac{1}{2}$. Pour une règle logique, IP3E vaut $\sqrt{1-\frac{1}{2^{n_a}}}$,

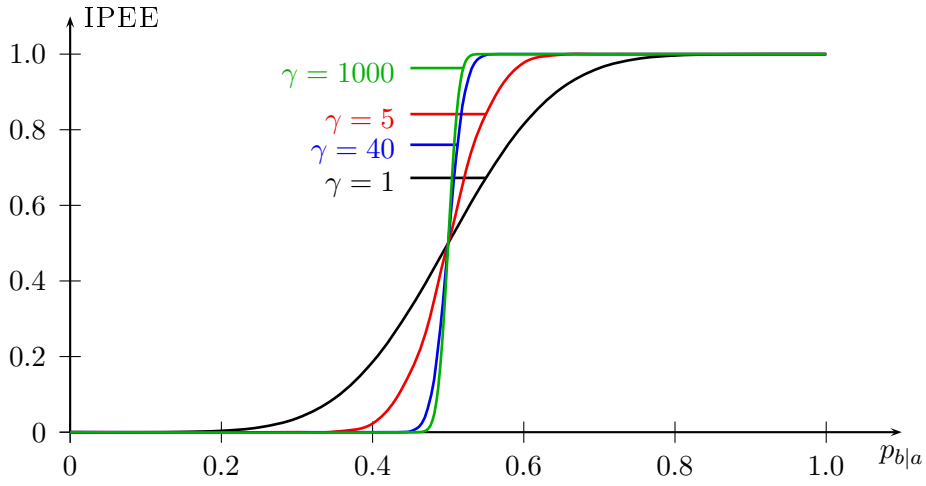


FIG. II.27 – Variations de IPEE en fonction de $p_{b|a}$

et est nulle lorsqu'il n'y a pas d'exemples. Sous hypothèse d'indépendance, les mêmes restrictions que pour IIE s'appliquent : IP3E est nulle, sauf lorsque $p_a < 1/2$ et $p_b > 1/2$.

Si l'on souhaite modifier sa définition afin qu'elle soit nulle à l'indépendance, on ne peut pas envisager de reprendre l'une des méthodes précédemment évoquées dans le cas d'IIE. En tronquant l'indice d'inclusion, on ne pourra annuler la valeur d'IP3E à l'indépendance, à cause du recentrage autour de la valeur 1/2. En effet, IP3E n'a pas de valeur constante à l'indépendance, ce qui ne sera pas corrigé en la multipliant par 1/2. On ne peut pas non plus appliquer celle ayant conduit à la construction de RIIE pour les mêmes raisons.

5 Synthèse

La table II.3 résume les expressions des mesures que nous venons de présenter. Leurs valeurs remarquables (valeurs minimales, maximales, à l'indépendance et à l'équilibre) sont présentées table II.4.

Parmi ces mesures, vingt seront étudiées formellement par la suite.

6 Application aux données "jouet"

Repartant de l'exemple jouet constitué au chapitre I, la table II.5 présente les évaluations numériques de chacune de ces règles pour vingt mesures. Les préordres générés par les mesures de qualité sont variés en nature (taille et nombre de groupes de règles), et en contenu (variabilité des ordonnancements des règles). Ils sont listés dans le tableau II.6. Trois règles y sont mises en valeur :

r_{18} est une règle qui est globalement bien évaluée par toutes les mesures. Il semble logique de globalement considérer qu'il s'agit d'une "bonne" règle.

Mesure	Référence	Définition absolue	Définition relative
SUP	[Agrawal et al., 1993]	$\frac{n_{ab}}{n}$	p_{ab}
CONF	[Agrawal et al., 1993]	$\frac{n_{ab}}{n_a}$	$\frac{p_{ab}}{p_a}$
R	[Pearson, 1896]	$\frac{nn_{ab}-n_an_b}{\sqrt{n_an_bn_{\bar{a}}n_{\bar{b}}}}$	$\frac{p_{ab}-p_ap_b}{\sqrt{p_ap_bp_{\bar{a}}p_{\bar{b}}}}$
LOE	[Loevinger, 1947]	$1-\frac{nn_{a\bar{b}}}{n_an_{\bar{b}}}$	$1-\frac{p_{a\bar{b}}}{p_ap_{\bar{b}}}$
CONFCEN		$\frac{nn_{ab}-n_an_b}{nn_a}$	$\frac{p_{ab}-p_a\times p_b}{p_a}$
CONV	[Brin et al., 1997b]	$\frac{n_an_{\bar{b}}}{nn_{a\bar{b}}}$	$\frac{p_ap_{\bar{b}}}{p_{a\bar{b}}}$
LAP	[Good, 1965]	$\frac{n_{ab}+1}{n_a+2}$	$\frac{p_{ab}+1/n}{p_a+2/n}$
FB	[Jeffreys, 1935]	$\frac{(n_a-n_{a\bar{b}})n_{\bar{b}}}{n_bn_{a\bar{b}}}$	$\frac{p_{ab}/p_b}{p_{a\bar{b}}/p_{\bar{b}}}$
GI	[Church and Hanks, 1990]	$\log(\frac{nn_{ab}}{n_an_b})$	$\log(\frac{p_{ab}}{p_ap_b})$
IQC	[Cohen, 1960]	$2\frac{nn_a-nn_{a\bar{b}}-n_an_b}{nn_a+nn_b-2n_an_b}$	$2\frac{p_{ab}-p_ap_b}{p_ap_{\bar{b}}+p_{\bar{a}}p_b}$
LIFT	[Brin et al., 1997a]	$\frac{nn_{ab}}{n_an_b}$	$\frac{p_{ab}}{p_ap_b}$
FUKUDA	[Fukuda et al., 1996]	$n_{ab}-\sigma_cn_a$	$np_a(\frac{p_{ab}}{pa}-\sigma_c)$
GAN	[Ganascia, 1987]	$2(n_{ab}/n_a-0.5)$	$2(p_{ab}/p_a-0.5)$
JAC	[Jaccard, 1908]	$\frac{n_{ab}}{n_{a\bar{b}}+n_b}$	$\frac{p_{ab}}{p_{a\bar{b}}+p_b}$
SEB	[Sebag and Schoenauer, 1988]	$\frac{n_{ab}}{n_{a\bar{b}}}$	$\frac{p_{ab}}{p_{a\bar{b}}}$
PS	[Piatetsky-Shapiro, 1991]	$n_{ab}-\frac{n_an_b}{n}$	$n(p_{ab}-p_ap_b)$
MOCO	[Azé and Kodratoff, 2002]	$\frac{n_{ab}-n_{a\bar{b}}}{n_b}$	$\frac{p_{ab}-p_{a\bar{b}}}{p_b}$
TEC		$\frac{n_{ab}-n_{a\bar{b}}}{n_{ab}}$	$\frac{p_{ab}-p_{a\bar{b}}}{p_{ab}}$
ZHANG	[Zhang, 2000]	$\frac{nn_{ab}-n_an_b}{\max\{n_{ab}n_{\bar{b}},n_bn_{a\bar{b}}\}}$	$\frac{p_{ab}-p_ap_b}{\max\{p_{ab}p_{\bar{b}},p_bp_{a\bar{b}}\}}$
-INDIMP	[Gras, 1979]	$\frac{nn_{ab}-n_an_b}{\sqrt{np_ap_{\bar{b}}}}$	$\sqrt{n}\frac{p_{ab}-p_ap_b}{\sqrt{p_ap_{\bar{b}}}}$
INTIMP	[Gras et al., 1996]	$P\Big[N(0,1)\geq \text{INDIMP}\Big]$	
IIE	[Gras et al., 2001]	$\sqrt{[(1-h_1(p_{a\bar{b}})^2)\times(1-h_2(p_{a\bar{b}})^2)]^{1/4}\times \text{INTIMP}}$	
RIIE	[Lallich et al., 2005b]	$\sqrt{[(1-h_1(p_{a\bar{b}})^2)\times(1-h_2(p_{a\bar{b}})^2)]^{1/4}\times\max\{2\,\text{INTIMP}-1;0\}}$	
TIIE	[Lallich et al., 2005b]	$\sqrt{[(1-h_1^*(p_{a\bar{b}})^2)\times(1-h_2^*(p_{a\bar{b}})^2)]^{1/4}\text{INTIMP}}$	
IPD	[Lerman and Azé, 2003]	$P\Big[\mathcal{N}(0,1)>\text{INDIMP}^{CR/\mathcal{B}}\Big]$	
IPEE	[Blanchard et al., 2005a]	$P\Big[N(0,1)\geq\frac{n_{a\bar{b}}-n_{ab}}{\sqrt{n_a}}\Big]$	$P\Big[N(0,1)\geq\sqrt{n}\frac{p_{a\bar{b}}-p_{ab}}{\sqrt{p_a}}\Big]$
IP3E	[Blanchard et al., 2005b]	$\sqrt{\frac{1}{2}\big\{[(1-h_1(p_{a\bar{b}})^2)\times(1-h_2(p_{a\bar{b}})^2)]^{1/4}+1\big\}\cdot\text{IPEE}}$	

avec :

$h_1(t)=-(1-\frac{t}{p_a})\log_2(1-\frac{t}{p_a})-\frac{t}{p_a}\log_2(\frac{t}{p_a})$ pour $t\in[0,p_a/2[$, $h_1(t)=1$ sinon

$h_2(t)=-(1-\frac{t}{p_{\bar{b}}})\log_2(1-\frac{t}{p_{\bar{b}}})-\frac{t}{p_{\bar{b}}}\log_2(\frac{t}{p_{\bar{b}}})$ pour $t\in[0,p_{\bar{b}}/2[$, $h_2(t)=1$ sinon

$h_1^*(t)=-(1-\frac{t}{p_a})\log_2(1-\frac{t}{p_a})-\frac{t}{p_a}\log_2(\frac{t}{p_a})$ pour $t\in[0,\min\{p_a/2;p_ap_{\bar{b}}\}[$, $h_1(t)=1$ sinon

$h_2^*(t)=-(1-\frac{t}{p_{\bar{b}}})\log_2(1-\frac{t}{p_{\bar{b}}})-\frac{t}{p_{\bar{b}}}\log_2(\frac{t}{p_{\bar{b}}})$ pour $t\in[0,min\{p_{\bar{b}}/2;p_ap_{\bar{b}}\}[$, $h_2(t)=1$ sinon

$\text{INDIMP}^{CR/\mathcal{B}}$ correspond à INDIMP, centré réduit étant donné une base de règle $\mathcal{B}$

TAB. II.3 – Mesures de qualité

mesure	minimum	indépendance	équilibre	maximum
Sup	0	$p_a p_b$	$p_a/2$	p_a
Conf	0	p_b	$1/2$	1
r	$-\sqrt{\frac{p_a p_b}{p_{\bar{a}} p_{\bar{b}}}}$	0	$\sqrt{\frac{p_a}{p_{\bar{a}} p_b p_{\bar{b}}}}(1/2 - p_b)$	$\sqrt{\frac{p_a p_{\bar{b}}}{p_{\bar{a}} p_b}}$
Loe	$-\frac{p_b}{p_{\bar{b}}}$	0	$\frac{1/2-p_b}{p_{\bar{b}}}$	1
ConfCen	$-p_b$	0	$1/2 - p_b$	$p_{\bar{b}}$
Conv	$p_{\bar{b}}$	1	$2p_{\bar{b}}$	$+\infty$
Lap	$\frac{1}{n_a+2}$	$\frac{n_a p_b+1}{n_a+2}$	$1/2$	$\frac{n_a+1}{n_a+2}$
FB	0	1	$p_{\bar{b}}/p_b$	$+\infty$
GI	$-\infty$	0	$-\log(2p_b)$	$\log \frac{n}{n_b}$
IQC	$-2\frac{p_a p_b}{p_a p_{\bar{b}} + p_{\bar{a}} p_b}$	0	$2\frac{p_a(1/2-p_b)}{p_a p_{\bar{b}} + p_{\bar{a}} p_b}$	$2\frac{p_a p_{\bar{b}}}{p_a p_{\bar{b}} + p_{\bar{a}} p_b}$
Lift	0	1	$1/p_b$	$\frac{n}{n_b}$
Fukuda	0	$n_a(p_b - \sigma_c)$	$n_a(1/2 - \sigma_c)$	$n_a(1 - \sigma_c)$
Gan	-1	$2p_b - 1$	0	1
Jac	0	$\frac{p_a p_b}{p_a p_{\bar{b}} + p_b}$	$\frac{p_a}{p_a + 2p_b}$	$\frac{p_a}{p_{a\bar{b}} + p_b}$
Seb	0	$p_b/p_{\bar{b}}$	1	$+\infty$
PS	$-n_a p_b$	0	$n_a(1/2 - p_b)$	$n a p_{\bar{b}}$
MoCo	$-\frac{p_a}{p_b}$	$p_a \frac{2p_b-1}{p_b}$	0	$\frac{p_a}{p_b}$
Tec	$-\infty$	$\frac{2p_b-1}{p_b}$	0	1
Zhang	-1	0	$\frac{1-2p_b}{\max\{p_b, p_{\bar{b}}\}}$	1
-IndImp	$-\sqrt{\frac{np_a}{p_{\bar{b}}}} p_b$	0	$\sqrt{\frac{np_a}{p_{\bar{b}}}}(1/2 - p_b)$	$\sqrt{n p_a p_{\bar{b}}}$
IntImp	0	$1/2$	$P\Big[N(0,1) \geq \sqrt{\frac{np_a}{p_{\bar{b}}}}(1/2 - p_b)\Big]$	1
IIE	0	$0^{\dagger}$	0	1
RIIE	0	0	0	1
TIIE	0	0	0	1
IPD ‡	$P\Big[N(0,1) \geq \frac{\sqrt{\frac{np_a}{p_{\bar{b}}}} p_b - m_{ii}}{e_{ii}}\Big]$	$P\Big[N(0,1) \geq -\frac{m_{ii}}{e_{ii}}\Big]$	$P\Big[N(0,1) \geq \frac{\sqrt{\frac{np_a}{p_{\bar{b}}}}(1/2-p_b) - m_{ii}}{e_{ii}}\Big]$	$P\Big[N(0,1) \geq -\frac{\sqrt{np_a p_{\bar{b}}} + m_{ii}}{e_{ii}}\Big]$
IPEE	0	$\sqrt{1 - \frac{1}{2^{n_a}}}$	$1/2$	1
IP3E	0	$0^{\dagger}$	$1/2$	$\sqrt{1 - \frac{1}{2^{n_a}}}$

† : IIE et IP3E ne sont pas nulles pour $p_a < 1/2$ et $p_b > 1/2$

‡ : m_{ii} et e_{ii} sont respectivement la moyenne et l'écart-type des valeurs prises par IndImp, sur la base de règle B considérée

Tab. II.4 – Valeurs extrêmes des mesures

r_{19} à l'inverse est constamment en fin de préordre. De manière similaire, il est logique de la considérer globalement comme mauvaise.

r_2 n'est ni toujours bonne, ni toujours mauvaise, et illustre notre problématique. Selon certaines mesures comme LIFT, GI ou encore R, elle est très bonne, mais selon d'autres mesures comme CONV ou MOCO elle est très mauvaise !

La variabilité des classements et des valeurs prises par les mesures nous a conduits à étudier l'agrégation de relations valuées afin de proposer un cadre prenant en compte cette diversité de situations [Barthélemy et al., 2006] (voir chapitre III).

7 Paramétrisation de mesures de qualité de règles d'association

La plupart des mesures de qualité objectives sont exprimées comme des fonctions de n, des fréquences marginales p_a et p_b, ainsi que de p_{ab}, ou $p_{a\bar{b}}$. On peut alors, puisque $p_{a\bar{b}} = p_a - p_{ab}$, les reformuler en fonction de n, des fréquences marginales p_a et p_b, et de la confiance.

Comme nous l'avons vu en section 2, plusieurs valeurs de référence peuvent sembler intéressantes pour l'évaluation d'une règle. Ainsi, la connaissance de la valeur d'une mesure à l'indépendance, ou à l'équilibre, permet de restreindre l'intervalle de choix qu'a un utilisateur afin de fixer son seuil minimal d'acceptation.

Si ce seuil est fixe, selon l'optique choisie (avoir des règles valides statistiquement, ou vraies à plus de 50%), la connaissance de la valeur prise par la mesure pourra guider de manière efficace l'utilisateur.

Or, nous avons noté par exemple que CONF n'ayant pas de valeur fixe à l'indépendance, un centrage additif ou multiplicatif sur p_b permettait de corriger ce biais. Si ce centrage est fait sur 1/2, comme dans le cas de GAN, on fixe la valeur à l'équilibre.

A notre avis, ces deux cas font référence à deux situations particulières. La première se fonde sur une logique de *ciblage*, alors que la seconde s'apparente plus à une logique de *prédiction* [Lallich et al., 2007].

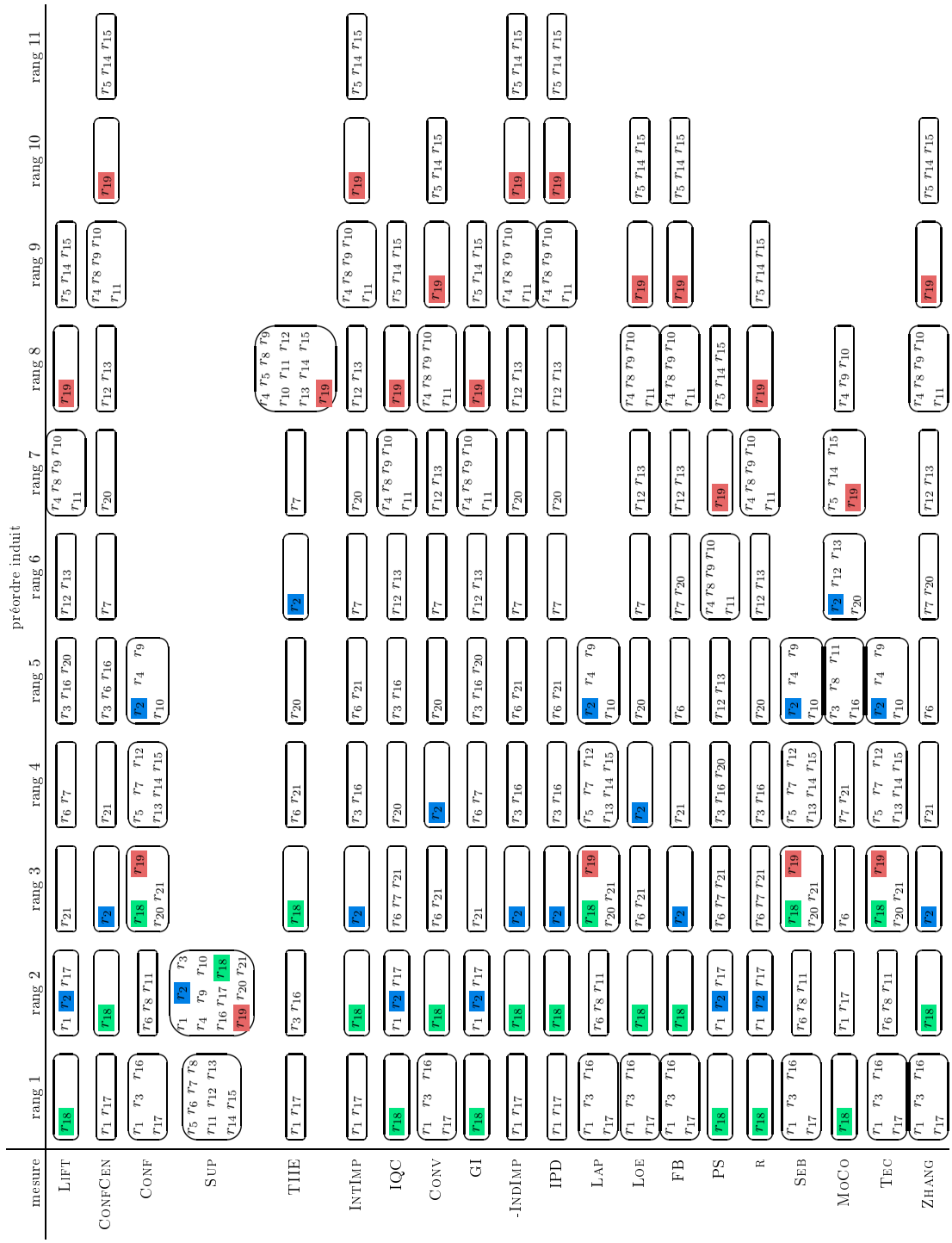
Dans [Lallich et al., 2005b] nous proposons une extension de ce concept de recentrage de la confiance, et proposons un ensemble de mesures généralisées. Ces mesures sont définies au moyen d'un paramètre θ, à fixer, et permettent d'adapter les mesures de qualité aux choix de référence de l'utilisateur. Prenant $\theta = p_b$, on se placera dans une logique de ciblage, et pour $\theta = 0.5$, dans une logique de prédiction.

7.1 Mesures descriptives

Pour généraliser les mesures de qualité listées en section 3 et 4, nous nous basons sur le constat suivant. Toutes les mesures que nous considérons prennent une valeur fixe soit à l'indépendance, soit à l'indétermination (voir table II.4), à l'exception de SUP, JAC et d'IPD. Dans le cas des mesures descriptives, il s'agit de plus d'une propriété exclusive : la

id	Lift	ConfCen	Conf	Sup	TIIE	IntImp	IQC	Conv	GI	-IndImp	IPD	Lap	LoE	FB	PS	R	Seb	MoCo	Tec	Zhang
1	2.00	0.50	1.00	0.30	0.94	0.89	0.60	$+\infty$	0.30	1.22	3.87	0.80	1.00	$+\infty$	1.50	0.65	$+\infty$	0.60	1.00	1.00
2	2.00	0.30	0.60	0.30	0.52	0.79	0.60	1.75	0.30	0.80	2.54	0.57	0.43	3.50	1.50	0.65	1.50	0.33	0.33	0.71
3	1.25	0.20	1.00	0.30	0.88	0.78	0.19	$+\infty$	0.10	0.77	2.45	0.80	1.00	$+\infty$	0.60	0.33	$+\infty$	0.38	1.00	1.00
4	1.00	0.00	0.60	0.30	0.00	0.50	0.00	1.00	0.00	0.00	0.00	0.57	0.00	1.00	0.00	0.00	1.50	0.17	0.33	0.00
5	0.83	−0.13	0.67	0.40	0.00	0.23	−0.36	0.60	−0.08	−0.73	−2.31	0.62	−0.67	0.50	−0.80	−0.41	2.00	0.25	0.50	−0.50
6	1.33	0.20	0.80	0.40	0.69	0.76	0.40	2.00	0.12	0.71	2.24	0.71	0.50	2.67	1.00	0.41	4.00	0.50	0.75	0.62
7	1.33	0.17	0.67	0.40	0.47	0.72	0.40	1.50	0.12	0.58	1.83	0.62	0.33	2.00	1.00	0.41	2.00	0.40	0.50	0.50
8	1.00	0.00	0.80	0.40	0.00	0.50	0.00	1.00	0.00	0.00	0.00	0.71	0.00	1.00	0.00	0.00	4.00	0.38	0.75	0.00
9	1.00	0.00	0.60	0.30	0.00	0.50	0.00	1.00	0.00	0.00	0.00	0.57	0.00	1.00	0.00	0.00	1.50	0.17	0.33	0.00
10	1.00	0.00	0.60	0.30	0.00	0.50	0.00	1.00	0.00	0.00	0.00	0.57	0.00	1.00	0.00	0.00	1.50	0.17	0.33	0.00
11	1.00	0.00	0.80	0.40	0.00	0.50	0.00	1.00	0.00	0.00	0.00	0.71	0.00	1.00	0.00	0.00	4.00	0.38	0.75	0.00
12	1.11	0.07	0.67	0.40	0.00	0.60	0.17	1.20	0.05	0.26	0.82	0.62	0.17	1.33	0.40	0.17	2.00	0.33	0.50	0.25
13	1.11	0.07	0.67	0.40	0.00	0.60	0.17	1.20	0.05	0.26	0.82	0.62	0.17	1.33	0.40	0.17	2.00	0.33	0.50	0.25
14	0.83	−0.13	0.67	0.40	0.00	0.23	−0.36	0.60	−0.08	−0.73	−2.31	0.62	−0.67	0.50	−0.80	−0.41	2.00	0.25	0.50	−0.50
15	0.83	−0.13	0.67	0.40	0.00	0.23	−0.36	0.60	−0.08	−0.73	−2.31	0.62	−0.67	0.50	−0.80	−0.41	2.00	0.25	0.50	−0.50
16	1.25	0.20	1.00	0.30	0.88	0.78	0.19	$+\infty$	0.10	0.77	2.45	0.80	1.00	$+\infty$	0.60	0.33	$+\infty$	0.38	1.00	1.00
17	2.00	0.50	1.00	0.30	0.94	0.89	0.60	$+\infty$	0.30	1.22	3.87	0.80	1.00	$+\infty$	1.50	0.65	$+\infty$	0.60	1.00	1.00
18	2.50	0.45	0.75	0.30	0.77	0.86	0.78	2.80	0.40	1.08	3.40	0.67	0.64	7.00	1.80	0.80	3.00	0.67	0.67	0.86
19	0.94	−0.05	0.75	0.30	0.00	0.41	−0.07	0.80	−0.03	−0.22	−0.71	0.67	−0.25	0.75	−0.20	−0.10	3.00	0.25	0.67	−0.25
20	1.25	0.15	0.75	0.30	0.63	0.68	0.23	1.60	0.10	0.47	1.50	0.67	0.38	2.00	0.60	0.25	3.00	0.33	0.67	0.50
21	1.50	0.25	0.75	0.30	0.69	0.76	0.40	2.00	0.18	0.71	2.24	0.67	0.50	3.00	1.00	0.41	3.00	0.40	0.67	0.67

TAB. II.5 – Evaluation des règles "jouet" par vingt mesures de qualité



Tab. II.6 – Préordres induits sur les règles "jouet" par vingt mesures de qualité

valeur est soit fixe à l'indépendance, soit fixe à l'indétermination, mais jamais simultanément à l'indépendance et à l'indétermination.

La généralisation que nous proposons vise à permettre de ne retenir que les règles telles que $1 \geq \textsc{Conf}(\mathtt{A} \rightarrow \mathtt{B}) \geq \theta \geq 0$, quelle que soit la mesure via la fixation d'un seuil nul ou constant, θ étant donné. Prenant par exemple $\theta = p_b$, on ne retiendra que les règles au delà de l'hypothèse d'indépendance, et prenant $\theta = 0.5$, on ne retiendra que les règles pour lesquelles $p_{ab} \geq p_{a\bar{b}}$, c'est-à-dire au delà de l'indétermination.

Nous allons ainsi chercher à exprimer les mesures comme une expression fonction de n, p_a et p_b, multipliée par l'une (ou un produit) des formes suivantes :

- $\textsc{Conf} - \theta$,
- $\frac{\textsc{Conf}}{\theta}$,
- $\frac{1-\textsc{Conf}}{\textsc{Conf}-\theta}$,
- $\frac{\textsc{Conf}-\theta}{1-\theta}$.

où $\theta = p_b$ lorsque la mesure prend une valeur fixe à l'indépendance, et $\theta = 1/2$ lorsque la mesure prend une valeur fixe à l'équilibre.

Le choix de ces quatre formes est lié aux réécritures possibles de l'inégalité nous intéressant :

$$
\begin{aligned}
1 \geq \textsc{Conf}(\mathtt{A} \rightarrow \mathtt{B}) \geq \theta \quad &\Leftrightarrow \quad 1 - \theta \geq \boxed{\textsc{Conf}(\mathtt{A} \rightarrow \mathtt{B}) - \theta \geq 0} \\
&\Leftrightarrow \quad 1 \geq \boxed{\tfrac{\textsc{Conf}(\mathtt{A} \rightarrow \mathtt{B}) - \theta}{1-\theta} \geq 0} \\
&\Leftrightarrow \quad +\infty \geq \boxed{\tfrac{1-\theta}{\textsc{Conf}(\mathtt{A} \rightarrow \mathtt{B}) - \theta} \geq 1} \\
&\Leftrightarrow \quad 1/\theta \geq \boxed{\tfrac{\textsc{Conf}(\mathtt{A} \rightarrow \mathtt{B})}{\theta} \geq 1}
\end{aligned}
$$

Les deux premières formes permettent de généraliser les mesures s'annulant à l'indépendance ou à l'équilibre, et les deux dernières formes permettent de généraliser celles prenant une valeur constante.

Toutes les mesures présentées précédemment ne se prêtent pas à cette réécriture. C'est le cas en particulier de Zhang et de Lap. Elles ne seront donc pas généralisables selon cette approche. Dans le cas de Lap, on peut toutefois proposer la généralisation recentrée sur 1/2 suivante : $\textsc{Lap}_{|\theta} = \frac{\textsc{Conf}-\theta}{1+2/n_a} + 1/2$. Toutefois, cette généralisation n'étant pas de nature similaire à celles que proposons, nous ne la retiendrons pas.

Etant donné une mesure μ, on écrira $\mu_{|\theta}$ sa version généralisée. Les formes généralisées des mesures sont les suivantes [Lallich et al., 2005b] :

- $\textsc{ConfCen}_{|\theta} = \textsc{Conf} - \theta$
- $\textsc{Gan}_{|\theta} = \frac{\textsc{Conf}-\theta}{1-\theta} = \frac{1}{1-\theta}\textsc{ConfCen}_{|\theta}$
- $\textsc{Loe}_{|\theta} = \frac{\textsc{Conf}-\theta}{1-\theta} = \textsc{Gan}_{|\theta}$
- $\textsc{Fukuda}_{|\theta} = np_a\,(\textsc{Conf} - \theta)$
- $\textsc{PS}_{|\theta} = np_a\,(\textsc{Conf} - \theta) = \textsc{Fukuda}_{|\theta}$
- $\textsc{Lift}_{|\theta} = \frac{\textsc{Conf}}{\theta}$
- $\textsc{GI}_{|\theta} = \log(\frac{\textsc{Conf}}{\theta}) = \log(\textsc{Lift}_{|\theta})$

- $\text{Conv}_{|\theta} = \frac{1-\theta}{1-\text{Conf}}$
- $\text{FB}_{|\theta} = \frac{\text{Conf}}{\theta} \times \frac{1-\theta}{1-\text{Conf}} = \text{Lift}_{|\theta} \times \text{Conv}_{|\theta}$
- $\text{Seb}_{|\theta} = \frac{\text{Conf}}{\theta} \times \frac{1-\theta}{1-\text{Conf}} = \text{FB}_{|\theta}$
- $\text{MoCo}_{|\theta} = \frac{\text{Conf}-\theta}{1-\theta} \times \frac{p_a}{p_b} = \text{Loe}_{|\theta} \times \frac{p_a}{p_b}$

Ces généralisations de mesures mènent dans certains cas à des formules identiques. Ainsi, on a :

- $\text{FB}_{|\theta=p_b} = \text{Seb}_{|\theta=0.5}$
- $\text{Gan}_{|\theta=0.5} = \text{Loe}_{|\theta=p_b}$
- $\text{Fukuda}_{|\theta=\sigma_c} = \text{PS}_{|\theta=p_b}$

7.2 Mesures statistiques et probabilistes

7.2.1 Rappels et complément

Comme vu en section 4, les mesures probabilistes sont construites à partir de la définition d'une mesure statistique, correspondant à la modélisation d'une catégorie de règles que l'on souhaite découvrir. Les mesures probabilistes s'expriment alors comme la probabilité d'observer une valeur de la mesure statistique au moins aussi grande que ce qui serait attendu sous hypothèse nulle, H_0. Classiquement, l'hypothèse nulle considérée est l'hypothèse d'indépendance entre les itemsets $\mathtt{A}$ et $\mathtt{B}$, et l'hypothèse alternative H_1 est l'hypothèse de dépendance positive entre $\mathtt{A}$ et $\mathtt{B}$.

[Lerman et al., 1981] ont proposé trois modélisations d'aléa afin de rendre compte de cette hypothèse d'indépendance. Nous reformulons ici en des termes équivalents ces trois modélisations, proposons une modélisation alternative au premier niveau d'aléa $\mathcal{N}_1$, que nous noterons $\mathcal{N}_{1'}$, et rappelons et introduisons ci-dessous les expressions de ces modélisations. Ces quatre modélisations sont présentées synthétiquement dans le tableau II.7.

On note N_{ab} la variable aléatoire générant n_{ab}, et H, B et Poi font respectivement références aux distributions hypergéométriques, binomiales et de Poisson.

Un niveau d'aléa : on raisonne à marges sont fixées, seules les fréquences conjointes sont aléatoires dans le tableau croisant $\mathtt{A}$ et $\mathtt{B}$, avec un seul degré de liberté.

La modélisation $\mathcal{N}_1$ proposée par [Lerman et al., 1981] considère la distribution des exemples, en prenant en compte les cellules intérieures de la table de contingence associée au couple ($\mathtt{A}$, $\mathtt{B}$), selon une méthodologie statistique classique.

- Sous H_0, N_{ab} suit la loi hypergéométrique $H(n, n_a, p_b)$. Tester H_0 revient alors à tester l'adéquation entre les confiances observées et les confiances théoriques de $\mathtt{A} \rightarrow \mathtt{B}$ et $\overline{\mathtt{A}} \rightarrow \mathtt{B}$, qui valent p_b à marges fixées.

Toujours à marges fixées, nous proposons dans [Lallich et al., 2005b] une approche alternative, $\mathcal{N}_{1'}$, qui ne prend en compte que la distribution des exemples entre $\mathtt{AB}$ et $\mathtt{A}\overline{\mathtt{B}}$.

- Dans ce cas, N_{ab} suit la loi binomiale $B(n_a, p_b)$. Tester H_0 revient alors à tester l'adéquation par rapport à la confiance observée et théorique de $\mathtt{A} \rightarrow \mathtt{B}$, p_b étant fixé.

Deux niveaux d'aléa : la modélisation $\mathcal{N}_2$ de [Lerman et al., 1981] correspond à la modélisation $\mathcal{N}_{1'}$, étendue à la situation où n_a est également aléatoire.

- Dans un premier temps, on suppose ainsi que N_a suit la loi binomiale $B(n, p_a)$.
- Dans un second temps, on exprime N_{ab}, conditionnellement à $N_a = n_a$, et N_{ab} suit alors la loi binomiale $B(n_a, p_b)$. N_{ab} suit ainsi la loi binomiale $B(n_a, p_a p_b)$.

Trois niveaux d'aléa : dans la modélisation $\mathcal{N}_3$ de [Lerman et al., 1981], les valeurs de n_a, puis n sont successivement rendues aléatoires (le raisonnement est identique dans la modélisation $\mathcal{N}_{1'}$).

- Au premier niveau d'aléa, N est supposé suivre la loi de Poisson $Poi(n)$.
- Dans un second niveau, on suppose que N_a suit la loi binomiale $B(n, p_a)$, conditionnellement à $N = n$.
- Enfin, conditionnellement à $N = n$ et $N_a = n_a$, on suppose que N_{ab} suit la loi binomiale $B(n_a, p_b)$. Dans ce cas, N_{ab} suit alors la loi de poisson $Poi(np_a p_b)$.

A partir de ces modélisations, on construit des indices statistiques $SI^{(i)}$, où i correspond à la modélisation retenue, en exprimant la loi de $N_{a\overline{b}}$ sous hypothèse nulle H_0. Selon la modélisation choisie, cette expression permet de centrer réduire $n_{a\overline{b}}$ selon l'espérance et l'écart-type attendu sous H_0 (notation ^{CR}), et on pose $SI^{(i)} = n^{CR}_{a\overline{b}}$.

Sous des conditions standard, les lois de distribution centrées réduites pouvant être approchées par la loi normale, on définit alors les mesures probabilistes $PI^{(i)}$, i faisant à nouveau référence à l'une des quatre modélisations retenues, comme étant le complément à 1 d'observer une forte valeur de l'indice statistique sous hypothèse nulle :

$$PI^{(i)} = P(N_{a\overline{b}} \geq n_{a\overline{b}}) = P(N(0,1) \geq SI^{(i)})$$

Le choix de la modélisation n'affecte pas l'espérance de $N_{a\overline{b}}$, qui vaut dans tous les cas de figure $n_a p_{\overline{b}}$, et seule la variance est affectée. Comme mentionné au préalable, [Gras, 1979] et [Lerman et al., 1981] préfèrent la troisième modélisation, qui est celle distinguant le plus les règles $\mathtt{A} \rightarrow \mathtt{B}$ et $\overline{\mathtt{B}} \rightarrow \overline{\mathtt{A}}$. A l'opposé, la première modélisation ne fait aucune distinction entre ces deux règles. On a ainsi IndImp $= SI^{(3)}$ et IntImp $= PI^{(3)}$.

Nous reprenons quelques mesures probabilistes déjà abordées afin de les resituer dans ce contexte.

Dans [Blanchard et al., 2005a], l'indice probabiliste d'écart à l'équilibre (IPEE) est construit sur un modèle similaire. De manière implicite, les auteurs se placent dans le cadre de la modélisation $\mathcal{N}_{1'}$, en posant $N_{a\overline{b}} \equiv B(n_a, 0.5)$, sous hypothèse d'indétermination, *i.e.* $N^{CR}_{a\overline{b}} = \frac{N_{a\overline{b}} - 0.5 n_a}{0.5\sqrt{n_a}}$. IPEE se définit alors comme : IPEE $= P\left[B(n_a, 0.5) > n_{a\overline{b}}\right]$.

Sous approximation normale, IPEE $= P\left[N(0,1) > \frac{n_{a\overline{b}} - 0.5 n_a}{0.5\sqrt{n_a}}\right]$, et vaut 0.5 à l'indétermination. Il correspond donc à la mesure probabiliste associée à $\mathcal{N}_{1'}$ (voir table II.7), dans laquelle p_b est remplacé par 0.5.

Bien que disposant de nombreuses bonnes propriétés (voir chapitre III), un défaut d'IntImp (défaut partagé avec les autres mesures probabilistes) est la perte de pouvoir discriminant. De par sa définition, IntImp évalue les règles significativement éloignées de l'indépendance entre 0.95 et 1. Dès lors que n devient important, ce qui est particulièrement

vrai dans un contexte de fouille de donnée, le moindre écart à l'indépendance sera signifiant, conduisant ainsi à des évaluations hautes et homogènes de toutes les dépendances positives, autour de 1. Afin de contre-balancer cette perte de pouvoir discriminant, deux approches ont été proposées et précédemment détaillées en 4.3 ([Gras et al., 2001]) et 4.5 ([Lerman and Azé, 2003]).

[Gras et al., 2001] proposent de pondérer les valeurs d'IntImp via un indice d'inclusion, basé sur l'entropie des expériences $\mathtt{B}|\mathtt{A}$ et $\overline{\mathtt{A}}|\overline{\mathtt{B}}$. Posant $H(X) = -p_x \log_2 p_x - p_{\overline{x}} \log_2 p_{\overline{x}}$ l'entropie associée un l'évènement X, la forme la plus générale de l'indice d'inclusion donnée dans [Blanchard et al., 2004] est $i(\mathtt{A} \subset \mathtt{B}) = \left[(1 - H^*(\mathtt{B}|\mathtt{A})^\alpha)\left(1 - H^*(\overline{\mathtt{A}}|\overline{\mathtt{B}})^\alpha\right)\right]^{\frac{1}{2\alpha}}$, avec $H^*(X) = H(X)$ si $p_x > 0.5$, et $H^*(X) = 1$ sinon.

Le paramètre α est à déterminer par l'utilisateur, et la valeur $\alpha = 2$ est conseillée par les auteurs si l'on souhaite que cet indice soit initialement tolérants aux contre-exemples. C'est donc cette valeur que nous considérerons par la suite.

[Gras et al., 2001] définissent alors l'intensité d'implication entropique comme étant : $\text{IIE} = [\text{IntImp} \cdot i(A \subset B)]^{\frac{1}{2}}$

Comme nous l'avons déjà mentionné, le recours à cet indice d'inclusion pose problème sous certaines conditions, puisque IIE n'a alors plus une valeur constante pour $p_a < 0.5$ et $p_b > 0.5$, et avons proposé dans [Lallich et al., 2005b] deux alternatives, TIIE et RIIE, permettant de corriger ce problème.

IPEE hérite également du problème de perte de pouvoir discriminant lorsque n croît. Les auteurs proposent ainsi de le moduler au moyen de l'indice d'inclusion, ce qui fait tout à fait sens, puisque ces deux mesures font référence à l'indétermination. Il n'y aura pas, comme dans le cas d'IIE de modification à apporter à l'indice d'inclusion pour que la version discriminante d'IPEE conserve sa valeur constante de 1/2 à l'indétermination.

Enfin, l'approche proposée par [Lerman and Azé, 2003] est de contextualiser la mesure probabiliste, en effectuant un recentrage et une réduction de IndImp sur l'ensemble des valeurs qu'il prend, pour la base de règles évaluée (notation $^{CR/\mathcal{B}}$). Ceci permet de définir l'indice probabiliste discriminant : $\text{IPD} = P\left[N(0,1) > \text{IndImp}^{CR/\mathcal{B}}\right]$.

7.2.2 Modélisations généralisées

En se basant sur une approche similaire à celle que nous avons mené afin de paramétriser les mesures descriptives, on peut généraliser les modèles $\mathcal{N}_1, \mathcal{N}_{1'}, \mathcal{N}_2$ et $\mathcal{N}_3$, et ainsi construire des versions paramétrisées des mesures de qualité statistiques et probabilistes. Nous allons donc chercher à exprimer les différents modèles en fonction de Conf, et de comparer sa valeur à un paramètre θ.

Nous généralisons chaque modélisation de la table II.7 de la manière suivante : sous H_0, la probabilité d'un exemple, conditionnellement à n_a, est θ. Ce choix se justifie naturellement de par le fait que Conf est la probabilité conditionnelle de n_{ab} sachant n_a. L'expression de N_{ab} est ainsi : $N_{ab} \equiv B(n_a, \theta)$.

Dans le cas des modélisations 1 et $1'$, les résultats sont immédiats, n_a étant fixé. Pour les

modélisations 2 et 3, l'expression de la loi de N_{ab} s'obtient à l'aide des fonctions génératrices des lois de probabilités :

- si $X \equiv B(m,p)$, sa fonction génératrice est $G(s) = E(s^X) = (1-p+ps)^m$,
- si $X \equiv Poi(\lambda)$, sa fonction génératrice est $G(s) = E(s^X) = e^{-\lambda(1-s)}$.

Dans la seconde modélisation, n est fixé, $N_a \equiv B(n,p_a)$ et $N_{ab}|(N_a = n_a) \equiv B(n_a,\theta)$. Puisque $G_{N_{ab}}(s) = E(s^{N_{ab}}) = E(E(s^{N_{ab}}/N_a)) = E\left((1-\theta+\theta s)^{N_a}\right)$, on a alors :

$$N_{ab} \equiv B(n,\theta p_a) \text{ et } N_{a\bar{b}} \equiv B(n,(1-\theta)p_a)$$

Dans la troisième modélisation, on a $N \equiv Poi(n)$, $N_a|(N=n) \equiv B(n,p_a)$, et $N_a|(N = n \text{ et } N_a = n_a) \equiv B(n_a,\theta)$.

Comme $G_{N_a}(s) = E(s^{N_a}) = E(E(s^{N_a}/N)) = E((1-p_a+p_as)^N) = e^{-np_a(1-s)}$, alors $N_a \equiv Poi(np_a)$.

De plus, la relation $G_{N_{ab}}(s) = E(s^{N_{ab}}) = E(E(s^{N_{ab}}/N_a)) = E((1-\theta+\theta s)^{N_a}) = e^{-n\theta p_a(1-s)}$ permet de déduire que :

$$N_{ab} \equiv Poi(n\theta p_a) \text{ et } N_{a\bar{b}} \equiv Poi(n(1-\theta)p_a)$$

A partir de l'expression de ces lois, on peut donc construire des versions paramétrisées des mesures statistiques et probabilistes selon l'approche classique consistant à évaluer l'écart entre le nombre de contre-exemples observé et celui attendu sous hypothèse nulle. La table II.7 résume les différentes mesures ainsi construites.

On note $SI^{(i)}_{|\theta} = n^{CR}_{a\bar{b}}$ l'indice statistique généralisé construit à partir de la modélisation i, et $PI^{(i)}_{|\theta} = P(N(0,1) \geq SI^{(i)}_{|\theta})$ la mesure probabiliste associée. A nouveau, la notation CR correspond ici à un recentrage et une réduction de $n_{a\bar{b}}$ sur l'espérance et l'écart-type de $N_{a\bar{b}}$.

7.3 Pouvoir discriminant des mesures généralisées

Tout comme leurs contreparties originales, les mesures généralisées perdent leur pouvoir discriminant lorsque n croît. On peut reprendre les approches proposées par [Gras et al., 2001] et [Lerman and Azé, 2003] afin de remédier à ce problème.

Selon l'approche contextuelle, on effectue un recentrage et une réduction de $\textsc{IndImp}_{|\theta}$ (ou l'une des autres expressions associées aux modélisations $\mathcal{N}_1$, $\mathcal{N}_{1'}$ ou $\mathcal{N}_2$) sur l'ensemble de ses valeurs prises, étant donné une base de règles $\mathcal{B}$. On définit ainsi l'indice probabiliste discriminant généralisé $\text{IPD}_{|\theta}$ comme suit :

$$\text{IPD}_{|\theta} = P\left(N(0,1) > \textsc{IndImp}^{CR/\mathcal{B}}_{|\theta}\right)$$

Notant $m_{ii_{|\theta}}$ et $e_{ii_{|\theta}}$ respectivement la moyenne et l'écart-type des valeurs prises par $\textsc{IndImp}_{|\theta}$ sur $\mathcal{B}$, $\textsc{IndImp}_{|\theta}$ prend alors la valeur constante $P\left(N(0,1) > -\frac{m_{ii_{|\theta}}}{e_{ii_{|\theta}}}\right)$ lorsque $\textsc{Conf} = \theta$. Cette grandeur dépend de la base $\mathcal{B}$ considérée, ainsi que du θ choisi.

Selon l'approche entropique, on définit l'intensité d'implication entropique généralisée ($\text{IIE}_{|\theta}$) et l'indice probabiliste d'écart à l'équilibre entropique généralisé ($\text{IP3E}_{|\theta}$) comme le

	Modélisations 1 et 1′	
Principe	1.1 : n_a est fixé, N_{ab} est aléatoire, et $\mathtt{A}$ et $\bar{\mathtt{A}}$ sont pris en compte	1'.1 : n_a est fixé, N_{ab} est aléatoire, et seul $\mathtt{A}$ est pris en compte
Loi de N_{ab} sous H_0	$H(n, n_a, p_b)$	$B(n_a, p_b)$
Loi de $N_{a\bar{b}}$ sous H_0	$H(n, n_a, p_{\bar{b}})$	$B(n_a, p_{\bar{b}})$
Indice statistique : $n^{CR}_{a\bar{b}}$	$SI^{(1)} = \frac{n_{a\bar{b}} - np_a p_{\bar{b}}}{\sqrt{np_a p_{\bar{a}} p_b p_{\bar{b}}}}$	$SI^{(1')} = \frac{n_{a\bar{b}} - np_a p_{\bar{b}}}{\sqrt{np_a p_b p_{\bar{b}}}}$
Mesure probabiliste : $P(N(0,1) > n^{CR}_{a\bar{b}})$	$PI^{(1)} = P(N(0,1) < \sqrt{n}\mathrm{R})$	$PI^{(1')}$

	Modélisation 2	**Modélisation 3**
Principe	2.1 : $N_a \equiv B(n, p_a)$ 2.2 : $N_{ab}\|N_a = n_a \equiv B(n_a, p_b)$	3.1 : $N \equiv P(n)$ 3.2 : $N_a\|N = n \equiv B(n, p_a)$ 3.3 : $N_{ab}\|N = n, N_a = n_a \equiv B(n_a, p_b)$
Loi de N_{ab} sous H_0	$B(n, n_a p_b)$	$Poi(np_a p_b)$
Loi de $N_{a\bar{b}}$ sous H_0	$B(n, n_a p_{\bar{b}})$	$Poi(np_a p_{\bar{b}})$
Indice statistique : $n^{CR}_{a\bar{b}}$	$SI^{(2)} = \frac{n_{a\bar{b}} - np_a p_{\bar{b}}}{\sqrt{np_a p_{\bar{b}}(1 - p_a p_{\bar{b}})}}$	$SI^{(3)} = \textsc{IndImp} = \frac{n_{a\bar{b}} - np_a p_{\bar{b}}}{\sqrt{np_a p_{\bar{b}}}}$
Mesure probabiliste : $P(N(0,1) > n^{CR}_{a\bar{b}})$	$PI^{(2)}$	$PI^{(3)} = \textsc{IntImp} = P(N(0,1) > \textsc{IndImp})$

Tab. II.7 – Indices statistiques et mesures probabilistes

	Modélisations 1 et 1′	
Principe	1.1 : n_a est fixé, N_{ab} est aléatoire, et $\mathtt{A}$ et $\bar{\mathtt{A}}$ sont pris en compte	1'.1 : n_a est fixé, N_{ab} est aléatoire, et seul $\mathtt{A}$ est pris en compte
Loi de N_{ab} sous H_0	$H(n, n_a, \theta)$	$B(n_a, \theta)$
Loi de $N_{a\bar{b}}$ sous H_0	$H(n, n_a, 1-\theta)$	$B(n_a, 1-\theta)$
Indice statistique : $n^{CR}_{a\bar{b}}$	$SI^{(1)}_{\|\theta} = \frac{n_{a\bar{b}} - np_a(1-\theta)}{\sqrt{np_a p_{\bar{a}}\theta(1-\theta)}}$	$SI^{(1')}_{\|\theta} = \frac{n_{a\bar{b}} - np_a(1-\theta)}{\sqrt{np_a\theta(1-\theta)}}$
Mesure probabiliste : $P(N(0,1) > n^{CR}_{a\bar{b}})$	$PI^{(1)}_{\|\theta} = P(N(0,1) < \sqrt{n}\mathrm{R})$	$PI^{(1')}_{\|\theta}$

	Modélisation 2	**Modélisation 3**
Principe	2.1 : $N_a \equiv B(n, p_a)$ 2.2 : $N_{ab}\|N_a = n_a \equiv B(n_a, p_b)$	3.1 : $N \equiv P(n)$ 3.2 : $N_a\|N = n \equiv B(n, p_a)$ 3.3 : $N_{ab}\|N = n, N_a = n_a \equiv B(n_a, p_b)$
Loi de N_{ab} sous H_0	$B(n, n_a\theta)$	$Poi(np_a\theta)$
Loi de $N_{a\bar{b}}$ sous H_0	$B(n, n_a(1-\theta))$	$Poi(np_a(1-\theta))$
Indice statistique : $n^{CR}_{a\bar{b}}$	$SI^{(2)}_{\|\theta} = \frac{n_{a\bar{b}} - np_a(1-\theta)}{\sqrt{np_a(1-\theta)(1-p_a(1-\theta))}}$	$SI^{(3)}_{\|\theta} = \textsc{IndImp}_{\|\theta} = \frac{n_{a\bar{b}} - np_a(1-\theta)}{\sqrt{np_a(1-\theta)}}$
Mesure probabiliste : $P(N(0,1) > n^{CR}_{a\bar{b}})$	$PI^{(2)}_{\|\theta}$	$PI^{(3)}_{\|\theta} = P(N(0,1) > \textsc{IndImp}_{\|\theta})$

Tab. II.8 – Indices statistiques et mesures probabilistes généralisés

produit de la mesure généralisée et d'un indice d'inclusion $i_{|\theta}$. Afin d'être cohérent, nous pensons qu'il est nécessaire que cet indice d'inclusion ait pour référence θ, et non plus 0.5.

Nous proposons ainsi la définition de l'indice d'inclusion généralisé construit comme suit. On définit $\widetilde{H}_{|\theta}(X)$ en substituant $\widetilde{p}_x$ à p_x dans la définition de $H(X)$, avec $\widetilde{p}_x = \frac{p_x}{2\theta}$ si $p_x \leq \theta$, $\widetilde{p}_x = \frac{p_x+1-2\theta}{2(1-\theta)}$ sinon (voir figure II.28). Ce changement permet de garantir que $\widetilde{H}_{|\theta}(\mathtt{B}|\mathtt{A}) = 1$ pour $\textsc{Conf} = \theta$.

Posant $\widetilde{H}^*_{|\theta}(X) = \widetilde{H}_{|\theta}(X)$ si $p_x > \theta$, $\widetilde{H}^*_{|\theta}(X) = 1$ sinon, on peut écrire :

$$i_{|\theta} = \sqrt[4]{\left(1 - \widetilde{H}^*_{|\theta}(\mathtt{B}|\mathtt{A})^2\right)\left(1 - \widetilde{H}^*_{|\theta}(\overline{\mathtt{A}}|\overline{\mathtt{B}})^2\right)}$$

Prenant $\theta = 1/2$, on retrouve bien l'indice d'implication. Toutefois, si la valeur de référence θ semble pertinente à utiliser afin de quantifier l'inclusion de $\mathtt{B}$ dans $\mathtt{A}$, elle ne fait pas de sens pour quantifier celle de $\overline{\mathtt{A}}$ dans $\overline{\mathtt{B}}$. En effet, puisque $p_{\overline{b}|\overline{a}} = 1 - \frac{p_a}{p_{\overline{b}}}(1 - \textsc{Conf})$, il peut sembler plus pertinent de considérer la valeur $1 - \frac{p_a}{p_{\overline{b}}}(1-\theta)$ comme référence dans la définition de $H(X)$, pour la partie de $i_{|\theta}$ faisant intervenir $\overline{\mathtt{A}}|\overline{\mathtt{B}}$. Dans ce cas, lorsque l'on souhaite faire référence à l'indépendance en prenant $\theta = p_b$, la valeur de référence est alors $p_{\overline{a}}$, et non pas p_b.

Il convient alors de distinguer les deux composantes de l'indice d'inclusion. On définit ainsi

$$\widetilde{H}^*_{1\,|\theta}(X) = \widetilde{H}^*_{|\theta}(X) = \begin{cases} 1 \quad \text{si } p_x \leq \theta \\ -\frac{1-p_x}{2(1-\theta)}\log_2\left(\frac{1-p_x}{2(1-\theta)}\right) - \left(1 - \frac{1-p_x}{2(1-\theta)}\right)\log_2\left(1 - \frac{1-p_x}{2(1-\theta)}\right) \quad \text{sinon} \end{cases}$$

qui quantifie l'inclusion de $\mathtt{A}$ dans $\mathtt{B}$, et

$$\widetilde{H}^*_{2\,|\theta}(X) = \begin{cases} 1 \quad \text{si } p_x \leq 1 - \frac{p_a}{p_{\overline{b}}}(1-\theta) \\ -\frac{p_{\overline{b}}}{p_a}\frac{1-p_x}{2(1-\theta)}\log_2\left(\frac{p_{\overline{b}}}{p_a}\frac{1-p_x}{2(1-\theta)}\right) - \left(1 - \frac{p_{\overline{b}}}{p_a}\frac{1-p_x}{2(1-\theta)}\right)\log_2\left(1 - \frac{p_{\overline{b}}}{p_a}\frac{1-p_x}{2(1-\theta)}\right) \quad \text{sinon} \end{cases}$$

qui quantifie l'inclusion de $\overline{\mathtt{B}}$ dans $\overline{\mathtt{A}}$. $\widetilde{H}^*_{2\,|\theta}(X)$ est défini comme $\widetilde{H}^*_{|\theta}(X)$, en prenant non plus θ, mais $1 - \frac{p_a}{p_{\overline{b}}}(1-\theta)$ comme référence, dans le changement de variable. Logiquement, on a bien $\widetilde{H}^*_{2\,|\theta}(X) = \widetilde{H}^*_{|1-\frac{p_a}{p_{\overline{b}}}(1-\theta)}(X)$.

L'indice d'inclusion que nous définissons est alors :

$$\hat{i}_{|\theta} = \sqrt[4]{\left(1 - \widetilde{H}^*_{1\,|\theta}(\mathtt{B}|\mathtt{A})^2\right)\left(1 - \widetilde{H}^*_{2\,|\theta}(\overline{\mathtt{A}}|\overline{\mathtt{B}})^2\right)} = \sqrt[4]{\left(1 - \widetilde{H}^*_{|\theta}(\mathtt{B}|\mathtt{A})^2\right)\left(1 - \widetilde{H}^*_{|1-\frac{p_a}{p_{\overline{b}}}(1-\theta)}(\overline{\mathtt{A}}|\overline{\mathtt{B}})^2\right)}$$

La définition de cet indice d'inclusion permet d'exprimer les mesures statistiques discriminantes généralisées suivantes :

$\text{IIE}_{|\theta} = \sqrt{\textsc{IntImp}_{|\theta} \times \hat{i}_{|\theta}}$

$\text{IP3E}_{|\theta} = \sqrt{\text{IPEE}_{|\theta} \times (\hat{i}_{|\theta} + 1)/2}$

La figure II.29 illustre le comportement de mesures probabilistes et leurs contreparties généralisées en fonction de CONF, pour $n = 1000$, $p_a = 0.05$ et $p_b = 0.10$. On observe nettement l'impact qu'a le paramètre θ afin de rendre les mesures généralisées plus ou moins discriminantes.

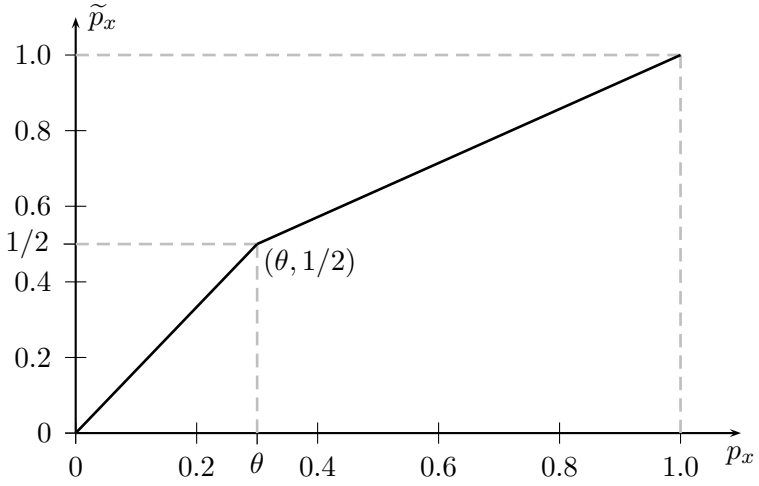


FIG. II.28 – Variations de $\widetilde{p}_x$ en fonction de p_x

Version originale		Version paramétrisée	
Mesure	Définition	Mesure	Définition
CONFCEN	$\text{CONF} - p_b$	$\text{CONFCEN}_{\|\theta}$	$\text{CONF} - \theta$
PS	$np_a(\text{CONF} - p_b)$	$\text{PS}_{\|\theta}$	$np_a(\text{CONF} - \theta)$
LOE	$\frac{\text{CONF}-p_b}{1-p_b}$	$\text{LOE}_{\|\theta}$	$\frac{\text{CONF}-\theta}{1-\theta}$
LIFT	$\frac{\text{CONF}}{p_b}$	$\text{LIFT}_{\|\theta}$	$\frac{\text{CONF}}{\theta}$
FB	$\frac{\text{CONF}}{p_b} \times \frac{1-p_b}{1-\text{CONF}}$	$\text{FB}_{\|\theta}$	$\frac{\text{CONF}}{\theta} \times \frac{1-\theta}{1-\text{CONF}}$
CONV	$\frac{1-p_b}{1-\text{CONF}}$	$\text{CONV}_{\|\theta}$	$\frac{1-\theta}{1-\text{CONF}}$
GI	$\log(\frac{\text{CONF}}{p_b})$	$\text{GI}_{\|\theta}$	$\log(\frac{\text{CONF}}{\theta})$
IQC	$2\frac{\text{CONF}-p_b}{p_{\bar{b}}+p_b\frac{p_{\bar{a}}}{p_a}}$	$\text{IQC}_{\|\theta}$	$2\frac{\text{CONF}-\theta}{p_{\bar{b}}+p_b\frac{p_{\bar{a}}}{p_a}}$
R	$\sqrt{\frac{p_a}{np_b(1-p_b)}}(\text{CONF} - p_b)$	$\text{R}_{\|\theta}$	$\sqrt{\frac{p_a}{np_b(1-p_b)}}(\text{CONF} - \theta)$

TAB. II.9 – Mesures descriptives paramétrisées, référence à l'indépendance

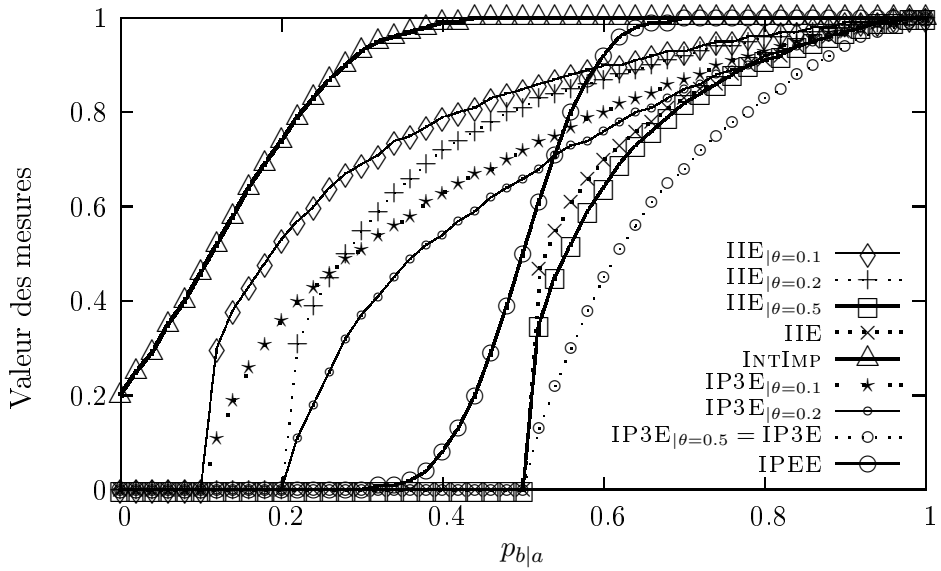


Fig. II.29 – Variations de mesures probabilistes et de leurs contreparties généralisées en fonction de $p_{b|a}$

Version originale		Version paramétrisée	
Mesure	Définition	Mesure	Définition
MoCo	$2 \times (\text{Conf} - 0.5) \times \frac{p_a}{p_b}$	$\text{MoCo}_{\|\theta}$	$2 \times (\text{Conf} - \theta) \times \frac{p_a}{p_b}$
Gan	$2 \times (\text{Conf} - 0.5)$	$\text{Gan}_{\|\theta}$	$2 \times (\text{Conf} - \theta)$
Seb	$\frac{\text{Conf}}{0.5} \times \frac{1-0.5}{1-\text{Conf}}$	$\text{Seb}_{\|\theta}$	$\frac{\text{Conf}}{\theta} \times \frac{1-\theta}{1-\text{Conf}}$
Tec	$\frac{2\times(\text{Conf}-0.5)}{\text{Conf}}$	$\text{Tec}_{\|\theta}$	$\frac{2\times(\text{Conf}-\theta)}{\text{Conf}}$

Tab. II.10 – Mesures descriptives paramétrisées, référence à l'indétermination

	Version originale		Version paramétrisée	
Modélisation	Mesure	Définition	Mesure	Définition
$\mathcal{N}_1$	$-\times-$	$\frac{n_{a\bar{b}} - np_a p_{\bar{b}}}{\sqrt{np_a p_{\bar{a}} p_b p_{\bar{b}}}}$	$-\times-$	$\frac{n_{a\bar{b}} - np_a(1-\theta)}{\sqrt{np_a p_{\bar{a}} \theta(1-\theta)}}$
$\mathcal{N}_{1'}$	$-\times-$	$\frac{n_{a\bar{b}} - np_a p_{\bar{b}}}{\sqrt{np_a p_b p_{\bar{b}}}}$	$-\times-$	$\frac{n_{a\bar{b}} - np_a(1-\theta)}{\sqrt{np_a \theta(1-\theta)}}$
$\mathcal{N}_2$	$-\times-$	$\frac{n_{a\bar{b}} - np_a p_{\bar{b}}}{\sqrt{np_a p_{\bar{b}}(1-p_a p_{\bar{b}})}}$	$-\times-$	$\frac{n_{a\bar{b}} - np_a(1-\theta)}{\sqrt{np_a(1-\theta)(1-p_a(1-\theta))}}$
$\mathcal{N}_3$	IndImp	$\frac{n_{a\bar{b}} - np_a p_{\bar{b}}}{\sqrt{np_a p_{\bar{b}}}}$	$\text{IndImp}_{\|\theta}$	$\frac{n_{a\bar{b}} - np_a(1-\theta)}{\sqrt{np_a(1-\theta)}}$

Tab. II.11 – Indices statistiques paramétrisés

	Version originale		Version paramétrisée	
Modélisation	Mesure	Définition	Mesure	Définition
$\mathcal{N}_1$	$-\times-$	$P\left[N(0,1) < \text{R}\right]$	$-\times-$	$P\left[N(0,1) > \frac{n_{a\bar{b}} - np_a(1-\theta)}{\sqrt{np_a p_{\overline{a}}\theta(1-\theta)}}\right]$
$\mathcal{N}_{1'}$	IPEE	$P\left[N(0,1) > \frac{n_{a\bar{b}} - 0.5n_a}{0.5\sqrt{n_a}}\right]$	$\text{IPEE}_{\|\theta}$	$P\left[N(0,1) > \frac{n_{a\bar{b}} - np_a(1-\theta)}{\sqrt{np_a\theta(1-\theta)}}\right]$
$\mathcal{N}_2$	$-\times-$	$-\times-$	$-\times-$	$P\left[N(0,1) > \frac{n_{a\bar{b}} - np_a(1-\theta)}{\sqrt{np_a(1-\theta)(1-p_a(1-\theta))}}\right]$
$\mathcal{N}_3$	IntImp	$P\left[N(0,1) > \text{IndImp}\right]$	$\text{IntImp}_{\|\theta}$	$P\left[N(0,1) > \text{IndImp}_{\|\theta}\right]$

TAB. II.12 – Mesures probabilistes paramétrisées

	Version originale		Version paramétrisée	
Mesure	Définition	Mesure	Définition	
IPD	$P\left[N(0,1) > \text{IndImp}^{CR/\mathcal{B}}\right]$	$\text{IPD}_{\|\theta}$	$P\left[N(0,1) > \text{IndImp}^{CR/\mathcal{B}}_{\|\theta}\right]$	
IIE	$\sqrt{\text{IntImp} \cdot i(A \subset B)}$	$\text{IIE}_{\|\theta}$	$\sqrt{\text{IntImp}_{\|\theta} \cdot \hat{i}(A \subset B)}$	
RIIE	$\sqrt{\text{IntImp}^* \cdot i(A \subset B)}$	$-\times-$	$-\times-$	
TIIE	$\sqrt{\text{IntImp} \cdot i_t(A \subset B)}$	$-\times-$	$-\times-$	

avec :

$\text{IntImp}^* = \max\{2\text{IntImp} - 1; 0\}$

$i_t(A \subset B) = \left[(1 - H_t^*(B/A)^\alpha)\left(1 - H_t^*(\overline{A}/\overline{B})^\alpha\right)\right]^{\frac{1}{2\alpha}}$

$H_t^*(B/A) = H(B/A)$ si $p_{b/a} > \max(0.5, p_b)$, $H_t^*(B/A) = 1$ sinon

$H_t^*(\overline{A}/\overline{B}) = H(\overline{A}/\overline{B})$ si $p_{\overline{a}/\overline{b}} > \max(0.5, p_{\overline{a}})$, $H_t^*(\overline{A}/\overline{B}) = 1$ sinon

TAB. II.13 – Mesures probabilistes discriminantes paramétrisées

Chapitre III

Deux approches "aide multicritère à la décision"

Nous avons vu au chapitre II qu'il existe un grand nombre de mesures. Chacune met en avant une catégorie de règles donnée, et toutes ne sont pas équivalentes, comme en atteste les rangements de l'exemple "jouet" du tableau II.6, ou encore les valeurs prises par les mesures pour des situations de référence, résumées en table II.4. Le chapitre IV, présentant des résultats expérimentaux, confirme ces divergences de comportement en situation.

L'expert métier (que nous noterons par la suite $E_{\mathcal{R}}$, pour "expert des règles") doit choisir quelle(s) mesure(s) utiliser pour filtrer un ensemble de règles (en fixant pour chaque mesure un seuil d'acceptation minimal). Selon la mesure utilisée et le seuil, les règles retenues ne seront pas les mêmes. Il se pose ainsi clairement le problème du choix d'une ou plusieurs mesures, et un mauvais choix peut conduire au rejet de règles potentiellement intéressantes, ou au contraire à la sélection de règles non souhaitées. La sélection des "bonnes" règles passe donc par l'utilisation de "bonnes" mesures [Lenca et al., 2003b]. Nous sommes confrontés à un problème de nature multicritère. Comme nous le verrons par la suite, les mesures peuvent être décrites selon différentes propriétés, les unes tantôt normatives, les autres tantôt subjectives, et il est alors pertinent de guider $E_{\mathcal{R}}$ dans son choix. L'assistance à l'expert métier est d'autant plus nécessaire qu'il n'existe pas de mesure possédant toutes les propriétés souhaitables.

L'aide multicritère à la décision (AMD par la suite) constitue un cadre particulièrement adapté à ce problème. En effet, l'AMD vise à guider, justifier ou rationaliser un choix parmi un ensemble d'actions $\mathcal{A} = \{a_1, \ldots, a_n\}$ possibles, ici par exemple le choix d'une mesure en vue d'une sélection pertinente de règles. Cette assistance se base sur la transformation des propriétés définies sur les mesures en critères de décision en intégrant les préférences des différents intervenants dans le processus. On distinguera ainsi l'expert $E_{\mathcal{A}}$ (expert de l'analyse multicritère, et de l'ECD) de l'expert $E_{\mathcal{R}}$ déjà mentionné. L'expert $E_{\mathcal{A}}$ exprimera ses préférences sur les propriétés que nous qualifions de normatives, l'expert $E_{\mathcal{R}}$ sur les propriétés subjectives. Nous nous plaçons, sauf mention contraire, en tant qu'expert $E_{\mathcal{A}}$.

Dans l'étude qui nous intéresse, étant donné deux alternatives quelconques, on demandera simplement à un critère d'être capable de nous indiquer quelle action est préférable sur

ce critère, ou si elles sont identiques. Dans le premier cas, on parlera de préférence, et dans le second d'indifférence. Chaque critère permet alors d'évaluer et de comparer entre elles les alternatives, et ils ne sont pas supposés avoir la même importance. Une troisième situation peut intervenir lorsque l'on prend en compte plusieurs critères, situation pour laquelle deux actions peuvent être déclarées incomparables.

L'AMD est donc particulièrement bien adaptée à notre problématique, à savoir sélectionner une ou plusieurs mesures, ou prendre en compte les divers avis exprimés par différentes mesures. Dans la première approche, l'ensemble des actions est constitué des mesures, alors que dans la seconde, il s'agit des règles.

De nombreuses méthodes AMD ont été proposées (voire [Roy, 1985], [Vincke, 1989], ou [Roy and Bouyssou, 1993] par exemple). Celles-ci s'adressent aux quatre problématiques de référence en AMD :

la problématique de choix : extraire de l'ensemble d'actions $\mathcal{A}$ un sous-ensemble restreint d'actions considérées comme les "meilleures" (voire figure III.1),

la problématique de tri : regrouper les actions de $\mathcal{A}$ en sous-groupes d'actions équivalentes (voire figure III.2),

la problématique de rangement : fournir un ordre sur les actions de $\mathcal{A}$, avec d'éventuels ex aequo, et des choix incomparables (voire figure III.3),

la problématique de synthèse : on souhaite synthétiser les éléments de décision (voire figure III.4).

Ces quatre problématiques ont été abordées dans nos travaux. Nous développerons particulièrement dans ce mémoire les problématiques de rangement et d'analyse. La classification formelle des mesures que nous proposons dans ce chapitre, et mettons en regard d'une classification expérimentale que nous présenterons dans le chapitre IV, s'inscrit dans une problématique de tri.

[Roy, 1985] liste trois types d'approches opérationnelles utilisées afin de répondre à ces problématiques :

1. l'approche par critère unique de synthèse, évacuant toute incomparabilité,
2. l'approche par surclassement de synthèse, acceptant l'incomparabilité,
3. l'approche par jugement local interactif avec itérations essai-erreur.

En étudiant les mesures selon une liste de propriétés pertinentes, et en systématisant leur évaluation sur ces dernières, nous apportons une première réponse à la problématique de synthèse.

Nous utiliserons dans ce mémoire la méthode de surclassement PROMETHEE ([Brans et al., 1984]), dont nous présentons les principaux aspects théoriques en section 1. Nous l'appliquons au rangement des mesures en section 2. Cette méthode est couplée à un outil de visualisation graphique, le plan GAIA, qui permet d'obtenir une synthèse de la situation. Cette facilité offerte au décideur contribue aux besoins d'analyse et est un élément qui a guidé notre choix sur l'utilisation de la méthode PROMETHEE.

Dans [Lenca et al., 2003b], nous abordons la problématique de choix avec la méthode ELECTRE I [Roy, 1968]. Nous avons conduit une étude similaire avec la version généralisée bipolaire d'ELECTRE proposée par [Bisdorff, 2004]. Enfin, la méthode TOMASO [Marichal

et al., 2005], qui réalise un rangement sur les alternatives, a également été appliquée. Ces différentes approches mènent à des résultats concordants, à savoir qu'elles ont toutes mis en avant les mêmes mesures.

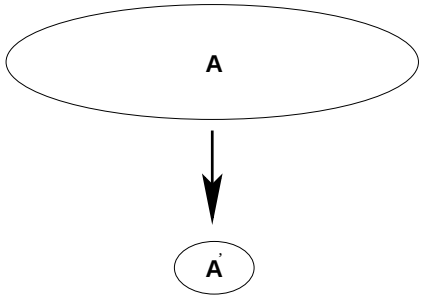


FIG. III.1 – Problématique de choix

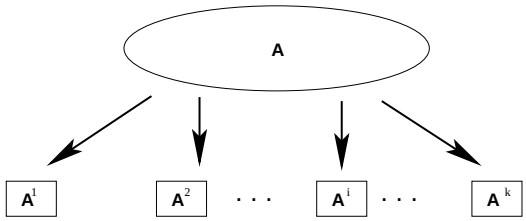


FIG. III.2 – Problématique de tri

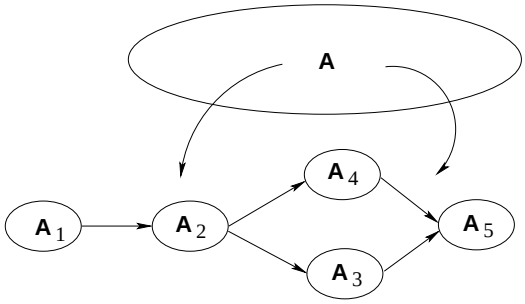


FIG. III.3 – Problématique de rangement

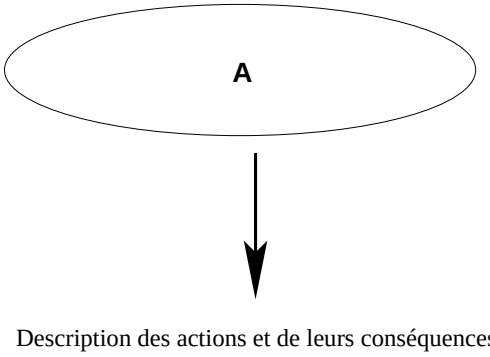


FIG. III.4 – Problématique de synthèse

Nous explorons également une voie alternative dans [Legrain, 2004, Barthélemy et al., 2006], cette fois fondée sur la construction d'un critère unique de synthèse. Ce critère est basé sur les évaluations des règles par les mesures. Il ne vise plus à choisir un groupe de mesures, mais à obtenir un résumé des évaluations d'un ensemble de règles par plusieurs mesures (éventuellement choisies par l'une des méthodes précédentes), et à prendre en compte des notions de suradditivité, de veto... Malheureusement, pour l'ensemble des opérateurs que nous avons étudiés, nous montrerons que les propriétés souhaitables ne peuvent pas être maintenues suite à l'opération d'agrégation des différentes informations prises en compte. L'étude d'autres opérateurs est une piste de recherche. Les aspects théoriques sont présentés en section 1, et une application à notre contexte est proposée en section 3.

1 Aspects théoriques

Chaque approche précédemment évoquée répond à un contexte donné. De manière générale :

- on aura recours à un critère unique de synthèse lorsque le modèle de préférence de l'utilisateur peut s'exprimer à travers une fonction unique. Ce type d'approche permet d'obtenir un rangement total des alternatives (il n'y a pas de choix incomparables, et la transitivité est préservée). Le résultat est donc aisément exploitable. En contrepartie, tous les critères doivent être commensurables. Il y a de plus une totale compensation entre les critères. Ce type d'approche est celui que l'on retrouve par exemple en théorie de l'utilité.
- on aura recours au surclassement lorsqu'un critère au moins n'est pas quantitatif, que les critères sont très hétérogènes et que leur codage sur une échelle commune est difficile, voire artificiel, ou encore que la compensation entre avantages et désavantages sur différents critères n'est pas justifiable. Ces approches permettent de tenir compte de seuils de préférence, ou de veto.
- les méthodes locales itératives sont utilisées lorsque l'espace des actions est grand, voire infini. Dès lors, on ne peut plus caractériser notre ensemble d'actions sur l'ensemble des critères de choix, et, à partir d'une solution aussi bonne que possible, on cherche s'il n'existe pas de meilleure solution dans son voisinage. Il s'agit généralement de méthodes complexes, et difficiles à mettre en œuvre en pratique.

Nous abordons les deux premières catégories de problèmes et en présentons ici les fondements théoriques. Leur application à notre contexte particulier de fouille de données sera développée en sections 2 et 3. D'autres travaux, comme [Lehn, 2000, Blanchard et al., 2003, Badjio and Poulet, 2006] adressent la troisième voie.

1.1 Approche PROMETHEE-GAIA

La méthode PROMETHEE-GAIA [Brans et al., 1984] a pour but de traiter des problèmes de nature multicritère en vue de construire un rangement des alternatives. Nous rappelons brièvement les principes de la méthode PROMETHEE dans cette section. Pour une présentation détaillée, nous renvoyons le lecteur vers [Brans and Mareschal, 1994, Brans and Mareschal, 2002].

PROMETHEE se base sur des comparaisons de paires d'alternatives sur les différents critères. Elle ne nécessite donc pas de critères commensurables. Cette comparaison sert à calculer un indice de préférence agrégé π pour chaque couple d'alternatives. Cet indice sert ensuite à calculer des flots de surclassement en vue de la création d'un ordre partiel sur les alternatives. L'ensemble des alternatives considérées ici est l'ensemble des m mesures $\mu_1, \ldots, \mu_m$, que nous notons désormais $\mathcal{M}$ (l'ensemble $\mathcal{A}$ précédent ayant un caractère générique).

De plus, une projection dans un plan particulier, le plan GAIA, permet de visualiser de manière très synthétique la structure du problème.

La construction des relations de surclassement est basée sur trois éléments.

Tout d'abord, nous définissons l'ensemble des dimensions servant à décrire le problème.

Il est constitué de k propriétés, décrivant le comportement des mesures de $\mathcal{M}$. On note $\mathcal{G} = \{g_1, \ldots, g_k\}$ cet ensemble. Un premier exemple de telle propriété est la tolérance à l'apparition des premiers contre-exemples à une règle. Un second est l'évaluation fixe ou non d'une règle logique par la mesure.

Dans un second temps, la prise en compte des préférences des différents intervenants impliqués dans le processus de décision nous conduit à transformer ces propriétés en un ensemble $\mathcal{F}$ de k critères. Par souci de simplification, nous noterons également ces critères g_i. On dira qu'une propriété est de nature normative lorsque les préférences sur cette propriété sont indépendantes du contexte applicatif et du décideur (elles dépendent de $E_{\mathcal{A}}$). Par exemple, comme nous l'avons déjà mentionné, il est préférable qu'une mesure prenne une valeur fixe pour les situations de référence à une mesure dont la valeur en ces points particuliers est dépendante de la règle évaluée. A l'opposé, lorsque le décideur a toute liberté d'exprimer ses préférences en fonction de ses objectifs, on parlera de propriétés subjectives (elles dépendent de $E_{\mathcal{R}}$). C'est le cas de la tolérance à l'apparition de contre-exemples à une règle.

La construction de l'ensemble de propriétés $\mathcal{G}$ permettant de prendre en compte, si possible, la totalité des points de vue exprimables constitue une première étape importante du processus d'AMD. Nous en avons étudié neuf de façon précise.

Une fois cet ensemble $\mathcal{G}$ défini, la traduction des préférences de $E_{\mathcal{R}}$ sur les propriétés de $\mathcal{G}$ en critères de décision est également une tâche non triviale, tout particulièrement lorsque la sémantique des propriétés est exprimée en langage naturel.

Les avis exprimés par $E_{\mathcal{A}}$ et $E_{\mathcal{R}}$ constituent le système de préférence $\mathcal{P}$ associé au problème.

Finalement, en croisant les informations contenues dans les éléments précédents, on obtient une matrice (m, k) de décision où chaque mesure en ligne est évaluée sur chaque critère en colonne.

On peut éventuellement ajouter une information supplémentaire relative à l'importance qu'accorderait le décideur aux différents critères à l'aide d'un système de poids.

La modélisation de préférence et la définition de critères a mobilisé de nombreux chercheurs en AMD [Vincke, 1989, Fodor and Roubens, 1994, Greco et al., 2002]. Etant donné la nature des préférences que nous allons construire sur les mesures, nous avons principalement utilisé lors de cette approche la notion de vrai critère :

Définition 7 (Critère de choix) *Un critère i est une fonction g_i définie sur $\mathcal{M}$ et prenant ses valeurs dans un ensemble ordonné $\mathbb{R}$, $g_i : \mathcal{M} \rightarrow \mathbb{R}$), et qui représente les préférences du décideur selon un point de vue.*

On dit que g_i est un :

- *vrai critère si la structure de préférence sous-jacente est une structure de pré-ordre total ("modèle traditionnel"),*
- *quasi-critère si la structure de préférence sous-jacente est une structure de quasi-ordre ("modèle à seuil"),*
- *pseudo-critère si la structure de préférence sous-jacente est une structure de pseudo-ordre.*

Le modèle de pseudo-critère est le plus général des trois ici présentés.

Les pseudo-critères g_i sont des fonctions à valeurs réelles définies sur $\mathcal{M}$, auxquelles on rattache deux seuils, le seuil p_i de préférence, et le seuil q_i d'indifférence. A partir de ces seuils, et étant donné deux objets a et b de $\mathcal{M}$, on définit les relations binaires de préférence P_i, de préférence faible Q_i, et d'indifférence I_i comme suit :

$$\begin{aligned} \mu_1 P_i \mu_2 &\Leftrightarrow g_i(\mu_1) - g_i(\mu_2) \geq p_i \\ \mu_1 Q_i \mu_2 &\Leftrightarrow q_i < g_i(\mu_1) - g_i(\mu_2) < p_i \\ \mu_1 I_i \mu_2 &\Leftrightarrow |g_i(\mu_1) - g_i(\mu_2)| \leq q_i \end{aligned}$$

La figure III.5 illustre les cinq situations de préférence possibles, en fonction de la valeur du pseudo-critère $g_i(\mu_2)$. On retrouve les deux autres types de critères :

- lorsque $p_j = q_j$, g_i est un quasi-critère,
- lorsque $p_j = q_j = 0$, g_i est un vrai-critère.

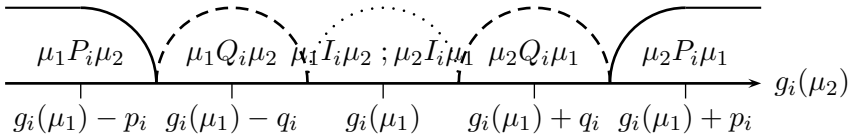


FIG. III.5 – Cinq situations de préférence construites sur $g_i(.)$

1.1.1 Comparaison paire par paire

La méthode PROMETHEE réalise des comparaisons par paire afin de déterminer le degré de préférence d'une alternative sur une autre. Pour une paire d'alternatives (μ_1, μ_2) de $\mathcal{M} \times \mathcal{M}$, pour un critère g_j, une petite (respectivement grande) différence d'évaluation sera représentée par un degré de préférence faible (respectivement élevé). Les auteurs de la méthode suggèrent de représenter ces préférences par des réels de l'intervalle $[0, 1]$ à l'aide d'une fonction de préférence P_j :

$$P_j(\mu_1, \mu_2) = P_j(g_j(\mu_1) - g_j(\mu_2)), \quad \text{où} \quad 0 \leq P_j(\mu_1, \mu_2) \leq 1,$$

le nombre d'arguments aux fonctions $P_j(.,.)$ et $P_j(.)$ permettant de distinguer deux familles de fonctions de préférences.

Dans le cas d'un critère à maximiser, la fonction de préférence $P_j(.)$ est une fonction croissante et continue par morceaux. C'est le cas de la fonction de préférence typique de la figure III.6.

L'approche PROMETHEE propose six types de fonctions de préférence qui nécessitent que l'utilisateur fixe entre 0 et 3 paramètres (seuils de préférence et d'indifférence). Suivant les

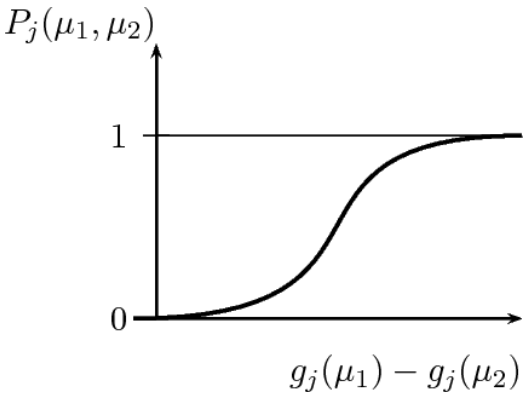


FIG. III.6 – Fonction de préférence typique

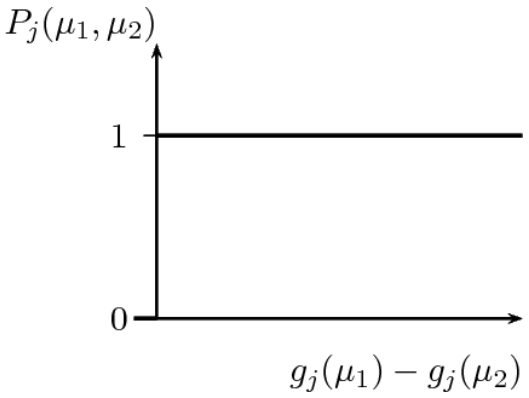


FIG. III.7 – Fonction de préférence usuelle

auteurs de cette méthode, ces six fonctions sont suffisantes pour l'analyse de la plupart des applications.

Vu la nature purement ordinale des critères que nous retiendrons lors de l'application de la méthode PROMETHEE à notre contexte, nous considérons qu'une supériorité stricte d'évaluation de μ_1 sur μ_2 doit être reflétée par la préférence de μ_1 sur μ_2 (modèle du vrai critère). Nous choisissons par conséquent d'utiliser la fonction de préférence usuelle, qui vaut 1 entre $]0, +\infty[$ et 0 partout ailleurs (figures III.6 et III.7).

1.1.2 Indice de préférence agrégé

L'étape suivante consiste à agréger les degrés de préférence relatifs à chaque critère en un indice de préférence agrégé, $\pi(\mu_1, \mu_2)$, qui exprime le degré avec lequel μ_1 est préféré à μ_2 sur l'ensemble des critères :

$$\pi(\mu_1, \mu_2) = \sum_{j=1}^{k} P_j(\mu_1, \mu_2) \cdot w_j$$

où w_j est le poids associé au critère g_j sous la contrainte $\sum_{j=1}^{k} w_j = 1$. On a $\pi(\mu, \mu) = 0$ et $0 \leq \pi(\mu_1, \mu_2) \leq 1, \forall \mu_1, \mu_2 \in \mathcal{M}$. Il est clair que pour une paire donnée (μ_1, μ_2) d'alternatives, un indice de préférence proche de 0 (respectivement 1) implique une préférence globale faible (respectivement forte) de μ_1 par rapport à μ_2.

1.1.3 Flots de surclassement

Le rangement des m alternatives de $\mathcal{M}$ est construit en comparant chacune des alternatives aux $m-1$ autres. Ceci se fait par l'intermédiaire de deux flots de surclassement définis comme suit :

- le flot de surclassement positif $\phi^+(\mu) = \frac{1}{m-1} \sum_{x \in \mathcal{M}} \pi(\mu, x)$
- le flot de surclassement négatif $\phi^-(\mu) = \frac{1}{m-1} \sum_{x \in \mathcal{M}} \pi(x, \mu)$

Le flot de surclassement positif exprime la force de l'alternative en question alors que le flot de surclassement négatif donne une indication sur sa faiblesse.

1.1.4 Les rangements PROMETHEE

Deux rangements différents peuvent être obtenus par la méthode PROMETHEE : un préordre partiel et un préordre total.

Définition 8 (Préordre) *Un préordre est une relation binaire réflexive, transitive. Lorsque cette relation est complète, on parle de préordre total.*

Le préordre partiel est l'intersection des rangements qui peuvent être déduits des flots de surclassement positifs et négatifs.

Il est construit de la façon suivante :

$$\begin{cases} \mu_1 P_p \mu_2 \iff \begin{cases} \phi^+(\mu_1) > \phi^+(\mu_2) \text{ et } \phi^-(\mu_1) < \phi^-(\mu_2) \\ \phi^+(\mu_1) = \phi^+(\mu_2) \text{ et } \phi^-(\mu_1) < \phi^-(\mu_2) \\ \phi^+(\mu_1) > \phi^+(\mu_2) \text{ et } \phi^-(\mu_1) = \phi^-(\mu_2) \end{cases} \\ \mu_1 I_p \mu_2 \iff \phi^+(\mu_1) = \phi^+(\mu_2) \text{ et } \phi^-(\mu_1) = \phi^-(\mu_2) \\ \mu_1 R_p \mu_2 \text{ sinon} \end{cases}$$

où P_p, I_p et R_p représentent respectivement la préférence, l'indifférence et l'incomparabilité.

Le préordre total est construit à partir du flot de surclassement net $\phi(a)$, avec $\phi(a) = \phi^+(a) - \phi^-(a)$:

$$\begin{cases} \mu_1 P_t \mu_2 \iff \phi(\mu_1) > \phi(\mu_2) \\ \mu_1 I_t \mu_2 \iff \phi(\mu_1) = \phi(\mu_2) \end{cases}$$

Bien sûr, on a $\mu_1 P_p \mu_2 \Rightarrow \mu_1 P_t \mu_2$ et $\mu_1 I_p \mu_2 \Rightarrow \mu_1 I_t \mu_2$.

1.1.5 Le plan GAIA

Après la détermination des rangements, la méthode PROMETHEE permet à l'utilisateur de visualiser les alternatives et les critères dans un plan commun appelé le plan GAIA. Il permet d'avoir une vue synthétique et claire du caractère conflictuel de certains critères, et de l'impact des poids sur la décision finale. Il s'agit d'une projection similaire à l'analyse en composantes principales des données. Cependant l'analyse du plan GAIA a l'avantage de permettre de visualiser le comportement du rangement final par rapport à différentes configurations de poids.

La construction du plan GAIA se base sur l'analyse de flots nets particuliers, relatifs à chacun des critères, en décomposant le flot de surclassement net.

Ainsi, on a :

$$\phi(\mu) = \phi^+(\mu) - \phi^-(\mu) = \frac{1}{n-1} \sum_{j=1}^{k} \sum_{x \in \mathcal{M}} [P_j(\mu, x) - P_j(x, \mu)] w_j$$

Par conséquent :

$$\phi(\mu) = \sum_{j=1}^{k} \phi_j(\mu) w_j \tag{III.1}$$

où $\phi_j(\mu) = \frac{1}{n-1} \sum_{x \in \mathcal{M}} [P_j(\mu, x) - P_j(x, \mu)]$ est le flot net unicritère. Chaque alternative μ peut ainsi être caractérisée par ses k flots unicritères $\alpha(\mu) = (\phi_1(\mu), \ldots, \phi_k(\mu))$ et être représentée dans un espace de dimension k dont les axes correspondent aux différents critères.

Le nuage de points correspondant est projeté dans un plan de dimension 2 de manière à présenter l'information de façon plus synthétique. Remarquons d'une part que dans l'espace de dimension k, le nuage de points représentant les alternatives est centré en l'origine, puisque $\sum_{\mu \in \mathcal{M}} \phi_j(\mu) = 0$. Cette projection est effectuée selon les deux premiers axes factoriels de l'analyse en composantes principales. D'autre part, dans le plan final, on représente aussi les projections des vecteurs unités des axes de coordonnées de l'espace original. Ces projections représentent alors les critères.

D'après la décomposition (III.1), le flot net d'une alternative est le produit scalaire entre le vecteur représentant son flot net unicritère et le vecteur des poids w. Ceci signifie en d'autres termes que le flot net de μ est aussi la projection de $\alpha(\mu)$ sur le vecteur w, dans l'espace de dimension k. Il en découle que la projection des $\alpha(x), \forall x \in \mathcal{M}$ sur w fournit le rangement total. Le vecteur w est donc un *axe de décision*. Sa projection dans le plan GAIA est appelée l'axe de décision PROMETHEE π.

Notons quelques points importants pour l'évaluation et l'analyse du plan GAIA :

- plus l'axe d'un critère est long dans le plan GAIA, plus ce critère sera discriminant pour les alternatives.
- des critères représentant des préférences similaires (respectivement opposées) sur l'ensemble d'alternatives sont représentés par des axes de direction similaire (respectivement opposée).
- des critères indépendants sont représentés par des axes orthogonaux.
- des alternatives qui ont de "bonnes" valeurs pour un critère particulier sont représentées par des points proches de l'axe de ce critère.
- des alternatives similaires sont voisines dans le plan GAIA.
- si l'axe π est long, il a un pouvoir de décision fort, et le décideur est invité à choisir des alternatives dans la direction et le sens de l'axe.
- si l'axe π est court, il a un pouvoir de décision faible. Dans ce cas, le vecteur w est quasi perpendiculaire au plan GAIA. Ceci signifie que pour cette configuration de poids, les critères sont en conflit, et un bon compromis proche de l'origine peut être choisi.

Le plan GAIA permet une exploration visuelle et interactive de la structure du problème.

L'utilisation de la méthode PROMETHEE, dont nous proposons une application en section 2, nous permettra de sélectionner une ou plusieurs mesures parmi les meilleures. Nous nous sommes donc naturellement posé la question de la prise en compte de l'avis exprimé par différentes mesures.

1.2 Approche par critère unique de synthèse

Une approche alternative au problème de choix de la mesure est la prise en compte des avis exprimés par différentes mesures. Dès lors, les objets sur lesquels se pose le problème multicritère ne sont plus les mêmes. Dans cette approche, les alternatives sont l'ensemble des règles $\mathcal{R}$. En exprimant un score évaluant la pertinence d'une règle, les mesures jouent le rôle de critères de choix. Se pose alors la question suivante : comment pouvons-nous effectuer l'agrégation de nos différentes mesures, tant leur nature et leur codomaine sont divers et hétérogènes ?

Trois voies se dégagent de la littérature :

Agrégation directe des évaluations des mesures Les mesures nous apportant une évaluation numérique de la qualité des règles, une première approche envisageable consiste à effectuer une agrégation complète, où l'on inclut toutes les performances dans une fonction d'utilité ou d'agrégation en leur affectant d'éventuels poids. Cette approche est entre autre étudiée par [Roy and Bouyssou, 1993], qui y font référence comme étant une "approche du critère unique de synthèse évacuant toute incomparabilité", ou par [Vincke, 1989] qui lui préfère le nom de "théorie de l'utilité multi-attribut". L'exemple le plus commun de ce type d'agrégation est la moyenne pondérée, classiquement utilisée pour évaluer un élève lors de son parcours scolaire.

Cette approche équivaut à la définition d'un critère global μ qui synthétise les différentes évaluations sur la règle r_i de $\mathcal{R}$ et que l'on définit de la façon suivante, avec g la fonction d'agrégation ou fonction d'utilité choisie :

$$\mu(r_i) = g(\mu_1(r_i), \ldots, \mu_m(r_i))$$

Cette méthode suppose entre autres que les jugements soient transitifs, ce qui est notre cas puisque chaque mesure induit un préordre total. Cependant, nous l'avons écartée car elle suppose aussi la commensurabilité des évaluations, du fait de l'hétérogénéité des mesures, et de leurs sémantiques variées.

Approche relationnelle pure L'approche relationnelle pure consiste en la modélisation de relations binaires de préférences individuelles à valeur dans {0,1} (les préférences étant observées pour toute paire de règles sur chacune des mesures, nous avons donc m relations de préférences individuelles par paire de règles) puis en leur agrégation ordinale. On obtient alors un profil Π de préférences individuelles. On se retrouve ainsi dans un contexte proche de celui de la théorie du choix social, qui étudie les différentes procédures de vote. Les votants sont ici les mesures, et les candidats les règles. On définit les préférences individuelles P_j entre deux règles r_i et r_k de $\mathcal{R}$ pour chaque mesure μ_j de la façon suivante :

$$r_i P_j r_k \quad \text{si} \quad \mu_j(r_i) \geq \mu_j(r_k)$$

Les relations P_j sont des préordres totaux, car transitives et complètes :
- la *transitivité* nécessite que $\forall r_i, r_k, r_l \in \mathcal{R}$, $r_i P_j r_k$ et $r_k P_j r_l \Rightarrow r_i P_j r_l$
- la *complétude* s'écrit $\forall r_i, r_k \in \mathcal{R}, r_i \neq r_k$, on a soit $r_i P_j r_k$ soit $r_k P_j r_i$

On désire désormais agréger ces m relations de préférence. Le théorème d'Arrow [Arrow, 1951, Kelly, 1991], valable dès que l'on souhaite agréger trois préférences ou plus, précise qu'aucune méthode d'agrégation ne peut satisfaire simultanément les conditions

d'universalité, de transitivité, d'unanimité, d'indépendance et de non dictature. Dans notre contexte ces différentes propriétés se traduisent comme suit :

- *universalité* : tout classement de règles est admissible.
- *transitivité* : l'agrégation doit fournir un préordre.
- *unanimité* : toute préférence unanime d'une règle sur une autre doit être conservée lors de l'agrégation.
- *indépendance* : le résultat de comparaison entre deux règles ne dépend que de leur position relative dans les préordres induits par les mesures.
- *non-dictature* : aucune mesure ne peut imposer son classement.

La construction de relations de préférences transitives peut révéler certains problèmes. Pour obtenir de telles relations, [Grabisch and Perny, 2003] font remarquer que l'on peut contourner ces difficultés en renonçant à l'une des conditions utilisées. La voie la plus raisonnable semble d'abandonner la condition d'indépendance d'Arrow. Dans notre contexte, les mesures étant contraintes (en particulier décroissantes avec le nombre de contre-exemples par hypothèse), on pourrait également envisager d'abandonner la condition d'universalité. Renoncer à la condition d'indépendance revient à accepter que le jugement final entre deux alternatives dépende non seulement des positions ou performances relatives de ces alternatives sur chacune des mesures, mais aussi de leurs positions ou performances relatives par rapport à d'autres alternatives. La méthode de Borda [Borda, 1781], qui consiste à calculer le rang de chaque alternative par critère pour ensuite construire un rang global, ne respecte pas cette condition d'indépendance.

Une autre approche est celle de Condorcet [Condorcet, 1785], qui compare les actions deux à deux, puis agrège les résultats. Cette méthode n'assure pas l'obtention d'un classement sur les actions. Enfin on peut aussi citer les méthodes lexicographiques qui privilégient les critères dont les poids sont les plus importants.

Utilisation des relations valuées Si l'approche relationnelle peut sembler intéressante à première vue, l'évaluation numérique qui nous est fournie par les mesures transmet certaines informations non négligeables, et il serait dommage de ne pas les prendre en compte. On peut aussi remarquer qu'une évaluation légèrement moins bonne d'une règle r_i sur une autre règle r_k peut être négligeable, ce qui peut traduire une indifférence ou une préférence faible alors qu'avec une modélisation nette des préférences nous aurions forcément $r_k P_j r_i$ et $r_i \neg P_j r_k$. Reprenant les règles r_1, r_2 et r_6 extraites du jeu de données jouet présenté au chapitre I, et observant les valeurs prises par Seb pour ces trois règles, on a $\text{Seb}(r_1) = +\infty$, $\text{Seb}(r_2) = 1.5$ et $\text{Seb}(r_6) = 2$. On interprète ces résultats comme "r_1 n'a pas de contre-exemple", "il y a une fois et demie plus d'exemples que de contre-exemples à r_2" et "il y a deux fois plus d'exemples que de contre-exemples à r_6". Clairement, r_1 est strictement préférée aux deux autres. Par contre, on préférera sans doute r_6 à r_2, mais pas avec la même intensité. Nous introduisons ainsi le concept de relations valuées, qui représentent le degré de crédibilité d'une affirmation. Elles sont définies sur $\mathcal{R} \times \mathcal{R}$ à l'aide de fonctions (continues ou non) à valeur dans $[0, 1]$. Elles sont utilisées dans les méthodes d'aide multicritère à la décision comme Promethee, précédemment évoquée, ou Electre III [Roy and Bouyssou, 1993]. [Bellman and Zadeh, 1970] ont été les premiers a établir un lien entre l'aide à la décision multicritère et la "théorie" des ensembles flous, l'idée de départ étant que l'ensemble des bonnes solutions associées à un critère correspond à un ensemble flou et que l'ensemble des solutions optimales est obtenu par la moyenne de l'intersection de ces ensembles flous.

C'est donc cette capacité qu'ont les relations valuées de traiter les catégories intermédiaires (différents degrés de préférence dans notre étude), tout en prenant en compte la nature hétérogène des évaluations, qui nous a décidé à favoriser l'approche valuée plutôt que l'approche "nette".

1.2.1 Présentation du processus d'agrégation

L'agrégation est un processus de combinaison de valeurs en une unique valeur, de façon à ce que le résultat final tienne compte d'une certaine façon des valeurs individuelles agrégées. Nous définissons une relation valuée de la façon suivante :

Définition 9 (Relation valuée) *Une relation valuée R est une fonction de l'ensemble des couples de règles vers* $[0,1]$ *:*

$$R : \mathcal{R} \times \mathcal{R} \rightarrow [0,1]$$

Le problème d'agrégation multicritère se modélise ici de la façon suivante : on souhaite définir une relation valuée R représentant une préférence globale. La valeur $R(r_i, r_k)$ affectée à toute paire d'alternatives (r_i, r_k), fonction des vecteurs de performances $(\mu_1(r_i), \ldots, \mu_m(r_i))$ et $(\mu_1(r_k), \ldots, \mu_m(r_k))$, s'écrit alors :

$$R(r_i, r_k) = f(\mu_1(r_i), \ldots, \mu_m(r_i), \mu_1(r_k), \ldots, \mu_m(r_k))$$

avec f une fonction définie sur $\mathbb{R}^{2m}$ à valeur dans $\mathbb{R}$ croissante des m premiers arguments et décroissante des m derniers arguments. f réalise une agrégation de performances mais aussi la comparaison des alternatives à l'aide de ces performances. On distingue deux modes opératoires différents en agrégation multicritère selon l'ordre de réalisation de ces deux étapes.

L'approche "agréger puis comparer" consiste à tout d'abord déterminer une agrégation des performances de chaque règle à l'aide d'un opérateur $\psi : \mathbb{R}^m \rightarrow \mathbb{R}$, puis à comparer les performances agrégées des deux règles r_i et r_k à l'aide d'une fonction $\phi_j : \mathbb{R}^2 \rightarrow \mathbb{R}$ pour $j = 1, \ldots, m$. Cette approche synthétise donc d'abord les évaluations par les mesures, puis les compare. Elle ne pourra ainsi tenir compte de l'écart entre deux règles, pour une mesure donnée, et ne répond donc pas à notre problématique.

Nous adoptons donc la seconde approche, "comparer puis agréger", qui consiste à tout d'abord comparer la performance de deux règles sur une mesure donnée, comparaison qui a un sens, puis d'agréger ces préférences partielles. R s'écrit alors :

$$R(r_i, r_k) = \psi[\phi_1(\mu_1(r_i), \mu_1(r_k)), \ldots, \phi_m(\mu_m(r_i), \mu_m(r_k))]$$

où les quantités du type $\phi_j(\mu_j(r_i), \mu_j(r_k))$ représentent des degrés de préférence entre deux règles.

Afin de construire notre relation valuée agrégée, il est donc nécessaire de se munir :

- d'une méthode de comparaison par mesure. Cette comparaison se fera au moyen de relations valuées spécifiques à chaque mesure. Le recours à ces relations valuées rend les grandeurs commensurables, et donc agrégeables via des opérateurs classiques,
- d'un opérateur d'agrégation synthétisant m relations valuées en une relation globale.

Au-delà de la simple définition de ces fonctions, nous allons en étudier quelques propriétés. Toute combinaison d'opérateurs n'aboutit pas à des résultats identiques. En particulier, rien ne garantit dans l'absolu que si une règle r_1 est préférée à une règle r_2 sur toutes les mesures, ceci soit encore le cas à l'issue du processus d'agrégation. L'étude des propriétés des opérateurs permet d'obtenir des éléments de réponse sur ces aspects. Ces études sont détaillées dans [Legrain, 2004], et nous ne reprenons ici que quelques aspects particuliers.

Notons également qu'à partir de la relation valuée agrégée, on peut se rapporter au cas binaire au moyen d'un seuil de coupure λ, en posant par exemple :

$$\mu_1 R \mu_2 \Leftrightarrow R(\mu_1, \mu_2) \geq \lambda$$

1.2.2 Définition de relations valuées

La première étape de l'agrégation multicritère consiste donc en la modélisation de relations de préférences individuelles définies sur $\mathcal{R}$ que nous noterons R_j pour $j = 1, \ldots, m$ et qui se formule :

$$R_j(r_i, r_k) = \phi_j(\mu_j(r_i), \mu_j(r_k))$$

Nous restreignons l'intervalle d'arrivée des fonctions ϕ_j et ψ à [0,1]. Souhaitant déterminer une préférence entre deux règles, nous proposons de l'évaluer en observant les écarts de performance sur chaque mesure, et écrivons la fonction ϕ_j – qui permet de définir la relation de préférence R_j sur la mesure μ_j – ainsi :

$$R_j(r_i, r_k) = h_j\left(\mu_j(r_i) - \mu_j(r_k)\right) = \phi\left(\mu_j(r_i), \mu_j(r_k)\right)$$

où h_j est une fonction croissante à valeurs dans $[0, 1]$. ϕ_j est alors bien croissante du premier argument, et décroissante du second.

1.2.3 Propriétés des relations valuées

Les relations valuées, tout comme les relations binaires usuelles, peuvent présenter différentes propriétés. Afin de pouvoir définir des propriétés sur des relations valuées qui soient significativement proches des propriétés "usuelles", il faut étendre les concepts usuels de conjonction/disjonction/négation. Ainsi on peut définir, par analogie avec le cas binaire, les termes de norme triangulaire (t-normes), conorme triangulaire (t-conormes) et de négation, dont on trouvera une étude plus approfondie dans [Fodor and Roubens, 1994].

Définition 10 (t-norme) *Une fonction $T : [0,1]^2 \to [0,1]$ est une t-norme si et seulement si elle vérifie :*

1. $T(1, x) = x, \forall x \in [0, 1]$
2. $T(x, y) = T(y, x), \forall x, y \in [0, 1]$ *(commutativité)*
3. $T(x, y) \leq T(u, v)$ *avec* $0 \leq x \leq u \leq 1, 0 \leq y \leq v \leq 1$ *(monotonie)*
4. $T(x, T(y, z)) = T(T(x, y), z), \forall x, y, z \in [0, 1]$ *(associativité)*

Définition 11 (t-conorme) *Une fonction $S : [0,1]^2 \to [0,1]$ est une t-conorme si et seulement si elle vérifie :*

1. $S(0,x) = x, \forall x \in [0,1]$
2. $S(x,y) = S(y,x), \forall x,y \in [0,1]$ *(commutativité)*
3. $S(x,y) \leq S(u,v)$ *avec* $0 \leq x \leq u \leq 1, 0 \leq y \leq v \leq 1$ *(monotonie)*
4. $S(x,S(y,z)) = S(S(x,y),z), \forall x,y,z \in [0,1]$ *(associativité)*

Définition 12 (négation) *Une fonction* $n : [0,1] \rightarrow [0,1]$ *est une négation si et seulement si elle vérifie :*

1. $n(0) = 1$ *et* $n(1) = 0$
2. n *est non croissante*

A toute t-norme T, on peut associer une t-conorme duale T^*, définie à l'aide de la négation standard $n(x) = 1 - x$:

$$T^*(x,y) = n(T(n(x),n(y)) = 1 - T(1-x, 1-y)$$

La donnée de ces trois opérateurs permet d'étendre des propriétés des relations binaires aux relations valuées, comme la transitivité, nécessaire pour garantir l'obtention d'un préordre[1] à l'issue de l'agrégation.

Parmi les t-normes/t-conormes fréquemment utilisées, citons :
- l'opérateur minimum comme t-norme [Zadeh, 1965], défini comme $T(x,y) = min(x,y)$. La t-conorme duale associée au min est $T^*(x,y) = max(x,y)$,
- l'opérateur maximum, dont la t-norme et sa t-conorme duale sont définies dans [Lukasiewicz, 1970] par $W(x,y) = max(x+y-1,0)$ et $W^*(x,y) = min(x+y,1)$.

Les propriétés de réflexivité, irréflexivité et symétrie s'étendent naturellement comme suit.

Définition 13 (réflexivité) *Une relation valuée est réflexive si et seulement si* $R(r_i,r_i) = 1, \forall r_i \in \mathcal{R}$.

Définition 14 (irréflexivité) R *est irréflexive si et seulement si* $R(r_i,r_i) = 0, \forall r_i \in \mathcal{R}$.

On remarquera qu'une relation de préférence réflexive suppose que toute règle est strictement préférée à elle-même alors que l'irréflexivité suppose qu'elle n'est absolument pas préférée à elle-même. L'irréflexivité semble plus appropriée à notre contexte étant donné la définition des R_j qui sont fonction croissante de la différence d'évaluation entre deux règles. Une relation valuée réflexive ne prend en compte que la non-préférence d'une règle sur une autre, alors qu'une relation irréflexive ne prend en compte que la préférence d'une règle sur une autre.

Définition 15 (symétrie) R *est symétrique si et seulement si* $R(r_i,r_k) = R(r_k,r_i), \forall r_i, r_k \in \mathcal{R}$.

Cette propriété n'est toutefois pas intéressante dans notre contexte, car elle ne permet pas d'orienter la préférence de la règle r_i sur la règle r_k.

[1] Rappelons qu'un préordre, dans le cas binaire, est défini comme une relation réflexive et transitive.

Afin de d'étendre d'autres propriétés aux relations valuées, nous allons nous appuyer sur le concept de t-normes.

La transitivité est une notion qui traduit l'idée que *le degré d'interaction entre deux éléments ne doit pas être plus faible que celui entre toute chaîne indirecte contenant d'autres éléments* [Fodor and Roubens, 1995]. La transitivité étant une condition primordiale dans toute problématique de classement elle a été le sujet de nombreuses études. Ainsi on distingue différentes formes de transitivité. On propose tout d'abord de la définir à l'aide de t-normes. Son expression générale est :

Définition 16 (T-transitivité) *Une relation valuée R sur $\mathcal{R}$ est T-transitive si et seulement si :*

$$T(R(r_i, r_k), R(r_k, r_l)) \leq R(r_i, r_l), \forall r_i, r_k, r_l \in \mathcal{R}$$

Afin d'obtenir une définition plus "parlante" de la transitivité, nous l'instancions à l'aide des t-normes précédemment évoquées. Nous obtenons alors les définitions suivantes :

Définition 17 (min-transitivité) *R est min-transitive (avec $T(x, y) = min(x, y)$) si et seulement si :*

$$R(r_i, r_l) \geq min(R(r_i, r_k), R(r_k, r_l)), \forall r_i, r_k, r_l \in \mathcal{R}$$

Cette propriété a l'avantage d'être très forte et traduit la volonté d'affecter une intensité de préférence entre deux règles qui ne doit pas être plus faible que celle entre toute chaîne contenant d'autres éléments.

Définition 18 (max-Δ transitivité) *R est max-Δ transitive (avec $T(x, y) = max(0, x + y - 1)$) si et seulement si :*

$$R(r_i, r_l) \geq max(0, R(r_i, r_k) + R(r_k, r_l) - 1), \forall r_i, r_k, r_l \in \mathcal{R}$$

Cette représentation correspond à l'inégalité triangulaire et n'a pas de signification sémantique particulière.

On peut également définir d'autres formes de transitivité, issues de la théorie des préférences stochastiques explicitées, voire par exemple [Tversky and Russo, 1969]. Ces relations prennent un sens quand les préférences sont décrites par des relations dites réciproques, *i.e.* qui vérifient $R + R^{-1} = 1$, c'est-à-dire

$$R(r_i, r_k) + R(r_k, r_i) = 1, \forall r_i, r_k \in \mathcal{R}.$$

Dans ce cas, $R(r_i, r_i) = 0.5$, soit $h_j(0) = 0.5$.

La valeur 0.5 à laquelle on compare $R(r_i, r_k)$ correspond au seuil à partir duquel on peut considérer que r_i est préféré à r_k. De par ce constat, on définit alors la faible transitivité comme suit.

Définition 19 (transitivité faible) *R est faiblement transitive si et seulement si :*

$$R(r_i, r_k) \geq 0.5 \text{ et } R(r_k, r_l) \geq 0.5 \Rightarrow R(r_i, r_l) \geq 0.5, \forall r_i, r_k, r_l \in \mathcal{R}$$

Alternativement, on peut à nouveau considérer les t-normes, il vient alors les définitions suivantes.

Définition 20 (transitivité modérée) *R est modérément transitive si et seulement si :*

$$R(r_i, r_k) \geq 0.5 \text{ et } R(r_k, r_l) \geq 0.5 \Rightarrow R(r_i, r_l) \geq min(R(r_i, r_k), R(r_k, r_l)), \forall r_i, r_k, r_l \in \mathcal{R}$$

Définition 21 (transitivité forte) *R est fortement transitive si et seulement si :*

$$R(r_i, r_k) \geq 0.5 \text{ et } R(r_k, r_l) \geq 0.5 \Rightarrow R(r_i, r_l) \geq max(R(r_i, r_k), R(r_k, r_l)), \forall r_i, r_k, r_l \in \mathcal{R}$$

Cette forme de transitivité est très forte car elle indique que dès qu'il y a préférence de r_i sur r_k et de r_k sur r_l, alors l'intensité de préférence de r_i sur r_l doit être plus forte que les deux degrés précédents.

Il existe certains liens entre ces différentes transitivités. On montre dans [Legrain, 2004] que :

$$\text{min-transitive} \Rightarrow \text{max-}\Delta \text{ transitive}$$

que :

$$\text{fortement transitive} \Rightarrow \text{modérément transitive} \Rightarrow \text{faiblement transitive},$$

et que :

$$\text{min-transitive} \Rightarrow \text{modérément transitive} \Rightarrow \text{faiblement transitive}.$$

En définissant de plus la notion de max-complétude et de min-antisymétrie (respectivement min-asymétrie), on pourrait alors introduire le concept de T-ordre partiel (respectivement T-ordre total), généralisant la notion d'ordre dans le cas valué. Toutefois, les contraintes imposées par ces notions ne seront pas vérifiées en pratique par les opérateurs que nous utiliserons. On pourra se référer à [Legrain, 2004] pour une définition de ces concepts.

1.2.4 Opérateurs d'agrégation

Définition 22 (Opérateur d'agrégation) *Un opérateur d'agrégation M est une fonction définie par :*

$$M : \bigcup_{m=1}^{\infty} [0,1]^m \rightarrow [0,1]$$

L'agrégation consiste en une évaluation numérique permettant de synthétiser les valeurs obtenues par nos relations valuées. Nous allons faire l'hypothèse que la valeur $R(x, y)$ de la relation agrégée ne dépend que du "vecteur" $(R_1(x, y), \ldots, R_m(x, y))$. Cette hypothèse n'est pas nécessairement réalisée puisque certaines procédures, notamment celles ayant recours à des fonctions de type score, comme la procédure de Borda, ne la respectent pas.

On peut affecter une pondération aux préférences individuelles permettant de traduire l'importance des différentes mesures. L'affectation de ces pondérations dépend des nécessités de l'expert. L'importance que celui-ci accorde aux critères sur les mesures éligibles permettra de définir un vecteur $W = (w_1, \ldots, w_m)$ représentatif où w_i est la pondération affectée à la mesure μ_i, avec $\sum_{i=1}^{m} w_i = 1$.

Un point important portant sur les opérateurs d'agrégation réside dans leur capacité à tenir compte de la majorité comme des minorités. En effet, de nombreux opérateurs d'agrégation ont tendance à sur-exagérer l'importance de certaines minorités au détriment de la majorité et vice versa. Ils ont été étudiés notamment par [Grabisch and Perny, 2003, Fodor and Roubens, 1994]. On distingue ainsi trois catégories d'opérateurs :

- les opérateurs **conjonctifs** dont le résultat d'agrégation est proche de 1 si les quantités à agréger sont toutes assez élevées. Les t-normes sont les opérateurs conjonctifs les plus fréquemment utilisés. De plus, grâce à la propriété d'associativité on peut étendre la définition d'une t-norme sur $[0,1]^2$ à une opération de $[0,1]^m \rightarrow [0,1]$ obtenue de façon récursive :
$$T(R_1, R_2, \ldots, R_n) = T(T(R_1, R_2, \ldots, R_{n-1}), R_n).$$
- les opérateurs **disjonctifs** qui affectent au résultat une valeur élevée dès que l'une des quantités à agréger est élevée et qui se modélisent à l'aide de t-conormes.
- les opérateurs de **compromis** vérifient $min_j R_j \leq M(R_1, \ldots, R_m) \leq max_j R_j$ et se situent donc entre ces deux précédentes classes d'opérateurs.

Les opérateurs de compromis sont ceux qui s'adaptent le mieux à notre problème. En effet, l'utilisation d'un opérateur disjonctif risque d'affecter une valeur proche de 1 après l'agrégation dès qu'une relation de préférence individuelle est élevée, traduisant ainsi une forte intensité de préférence globale. Pourtant, ce n'est pas parce qu'on observe une forte préférence d'une règle sur une autre pour une mesure considérée, surtout si elle est considérablement moins préférée pour les autres mesures, qu'on peut affirmer qu'elle est globalement préférée à l'autre. Nous pouvons retourner cette observation dans le cas d'un opérateur conjonctif qui affecterait une préférence globale élevée si et seulement si toutes les préférences individuelles sont élevées.

Nous définissons ici la classe d'opérateurs d'agrégation la plus simple : la moyenne généralisée. D'autres opérateurs (opérateurs OWA, intégrales floues,...) ont été étudiés dans ce contexte dans [Legrain, 2004].

Définition 23 (Moyennes généralisées) *La moyenne généralisée pondérée se définit comme :*
$$M(R_1, \ldots, R_m) = f^{-1}[\sum_{j=1}^{m} w_j f(R_j)]$$
où f est une fonction continue inversible et w_j correspond au poids de R_j et vérifie $\sum_{j=1}^{m} w_j = 1$.

Cet opérateur regroupe notamment :

- moyenne arithmétique pondérée avec $f(x) = x$:
$$MA(R_1, \ldots, R_m) = \sum_{i=1}^{m} w_j R_j$$
- moyenne géométrique pondérée avec $f(x) = log(x)$:
$$MG(R_1, \ldots, R_m) = [\prod_{i=1}^{m} R_j^{w_j}]$$

- moyenne d'ordre α ou root-power mean avec $f(x) = x^\alpha$:

$$M_\alpha(R_1, \ldots, R_m) = [\sum_{j=1}^{m} w_j R_j^\alpha]^{\frac{1}{\alpha}}$$

- moyenne harmonique pondérée avec $f(x) = \frac{1}{x}$:

$$MH(R_1, \ldots, R_m) = [\sum_{j=1}^{m} w_j R_j^{-1}]^{-1}$$

Les moyennes généralisées ont la particularité d'être des opérateurs continus, strictement monotones et idempotents.

Définition 24 (Stricte monotonie) *Un opérateur d'agrégation M est dit strictement monotone si*

$$R'_i > R_i \Rightarrow M(R_1, \ldots, R'_i, \ldots, R_m) > M(R_1, \ldots, R_i, \ldots, R_m), \forall i \in [\![1, m]\!]$$

Définition 25 (Idempotence) *Un opérateur d'agrégation M est dit idempotent s'il satisfait la condition d'unanimité, i.e. :*

$$M(R_j, \ldots, R_j) = R_j$$

Nous nous intéressons principalement à la conservation de la propriété de transitivité, dans la mesure où nous souhaitons définir un préordre sur l'ensemble des règles après l'agrégation. Le tableau III.1, issu de [Peneva and Popchev, 2003], permet de synthétiser des résultats obtenus.

En observant ce tableau, on remarque que la transitivité est toujours "affaiblie" après l'agrégation : la propriété de min-transitivité que nous aurions aimé conserver ne peut être conservée. En effet, hors la max-Δ-transitivité, les propriétés de transitivité ne sont guère conservées par les opérateurs d'agrégation. Un théorème dû à [Leclerc, 1984] caractérise les fonctions d'agrégation min-transitives. Celles-ci sont de la forme $\wedge_{1 \leq j \leq m} \varphi_j(R_j(r_i, r_k))$ où φ_j est une fonction monotone. Ces opérateurs sont des "oligarchies"[2], ne répondant donc certainement pas à notre souhait de prendre en compte les divers avis exprimés.

De plus, la réflexivité des relations valuées est nécessaire. On peut donc d'ores et déjà s'attendre à devoir faire des compromis en pratique, l'ensemble des propriétés que nous souhaitions retenir ne pouvant être vérifié simultanément.

La mise en œuvre de cette approche est présentée en section 3.

2 Aide à la sélection de mesures de qualité

Nous mettons en œuvre la méthode de surclassement PROMETHEE présentée en section 1. Une des premières étapes de cette mise en œuvre consiste à définir l'ensemble des critères nécessaires à l'étude des mesures de qualité.

[2]Terme formé du grec *oligos*, peu nombreux et *arkhê*, commandement.

Opérateur	Propriétés vérifiées par les R_j	Propriétés déduites de R
Moyenne arithmétique pondérée	max-Δ transitif + réflexif	faiblement transitif + réflexif
Moyenne géométrique pondérée	réflexif + min-transitif	réflexif + max-Δ transitif
	réflexif + min-asymétrie + min-transitif	réflexif + min-asymétrie + max-Δ transitif
Moyenne harmonique	réflexif + min-transitif	réflexif + max-Δ transitif
Moyenne géométrique	réflexif + min-transitif	réflexif + max-Δ transitif
Moyenne arithmétique	réflexif + min-transitif	réflexif + max-Δ transitif

TAB. III.1 – Conservation des propriétés par les opérateurs d'agrégation

Ainsi, un premier objectif a guidé notre travail, à savoir déterminer de façon la plus exhaustive possible les diverses propriétés qu'un utilisateur pourrait souhaiter qu'une mesure possède. Nous espérons ainsi apporter une contribution originale en ECD en présentant une étude transversale de vingt mesures de qualité, sur neuf propriétés. Parmi les propriétés retenues, certaines sont issues de la littérature, et nous en avons proposé de nouvelles. Par ailleurs, très peu d'articles de la littérature ont systématisé l'étude d'un nombre important de mesures sur un nombre important de propriétés. A notre connaissance, seuls [Hilderman and Hamilton, 2001b], [Hilderman and Hamilton, 2003] et [Tan et al., 2002] ont abordé le sujet dans des dimensions comparables. Notre travail débouche sur la construction d'une matrice évaluant chaque mesure sur neuf propriétés. La prise en compte de préférences utilisateur sur chaque propriété étudiée permet alors de mettre en œuvre un processus d'aide multicritère à la décision (AMD) afin de guider le choix d'une mesure.

Nous nous positionnons un premier temps dans un contexte purement analytique, et n'exprimons aucune préférence sur les propriétés que nous étudions. Dans un second temps, en section 2.9, nous nous placerons dans un contexte d'AMD, en tant qu'expert de la fouille de donnée, $E_{\mathcal{A}}$, et illustrerons comment un tel processus permet d'assister l'expert des données, $E_{\mathcal{R}}$, afin de sélectionner une mesure de qualité pertinente.

Nous présentons successivement les neuf propriétés retenues. Pour chaque propriété, nous proposons un ensemble de valeurs possibles. Bien que, comme nous l'avons déjà vu précédemment, certaines propriétés puissent être fortement souhaitées par l'utilisateur, nous rappelons que n'imposons pas ici d'ordre sur les modalités de ces propriétés, la prise en compte d'un système de préférences est développée en section 2.9.

Parmi les mesures présentées au chapitre II, nous avons retenu vingt mesures pour illustrer notre propos. Chaque mesure sera évaluée selon les neuf propriétés. Ces évaluations sont résumées dans le tableau III.26.

Nous avons bien évidemment retenu SUP et CONF, de par leur rôle primordial en extraction de règles d'association. Nous avons par ailleurs étudié les propriétés des mesures descriptives R, LOE, CONFCEN, CONV, LAP, FB, GI, IQC, LIFT, SEB, MOCO, PS, TEC et ZHANG, ainsi que celles de -INDIMP, IPD, INTIMP et TIIE pour les indices statistiques et les mesures probabilistes.

Bien évidemment, la démarche proposée peut être étendue à d'autres mesures et à d'autres propriétés.

Les propriétés que nous avons étudiées sont les suivantes[3] :

[3]Nous tenons à remercier les membres du Groupe de travail GAFOQUALITÉ (animé par Fabrice Guillet,

- traitement orienté de la prémisse et de la conclusion [Freitas, 1999],
- décroissance selon la taille du conséquent [Piatetsky-Shapiro, 1991],
- évaluation de situations de références :
  - valeur pour une règle logique [Lenca et al., 2003a],
  - valeur à l'indépendance [Piatetsky-Shapiro, 1991],
  - valeur à l'indétermination [Blanchard et al., 2005a],
- tolérance aux premiers contre-exemples [Gras et al., 2001],
- prise en compte du nombre de cas [Lallich, 2002, Gras et al., 2004],
- facilité à fixer un seuil d'acceptation de la règle [Lenca et al., 2003b, Lallich and Teytaud, 2004],
- intelligibilité de la mesure [Lenca et al., 2003a].

Certaines propriétés sont augmentées de discussions autour de travaux en cours et permettant de compléter l'analyse faite de l'espace des mesures.

2.1 Traitement orienté de la prémisse et de la conclusion

Les règles d'association sont des connaissances orientées. En effet, dans la règle $\mathtt{A} \rightarrow \mathtt{B}$, la donnée de l'itemset $\mathtt{A}$ doit permettre d'augmenter la connaissance que l'on a de l'itemset $\mathtt{B}$ *a priori* (c'est-à-dire p_b). Les rôles respectifs de $\mathtt{A}$ et $\mathtt{B}$ sont donc bien distincts, et la règle $\mathtt{A} \rightarrow \mathtt{B}$ n'a pas le même sens que la règle $\mathtt{B} \rightarrow \mathtt{A}$. Ainsi, on peut souhaiter qu'une mesure de qualité μ évalue différemment ces deux règles *i.e.* que μ soit orientée selon $\mathtt{A}$ et $\mathtt{B}$ [Freitas, 1999].

On dira donc qu'une mesure μ est orientée (propriété g_1), s'il existe au moins un couple d'itemsets ($\mathtt{A}$,$\mathtt{B}$) tels que $\mu(\mathtt{A} \rightarrow \mathtt{B}) \neq \mu(\mathtt{B} \rightarrow \mathtt{A})$.

On notera alors $g_1(\mu) = \mathtt{ori}$ lorsqu'une mesure est orientée, et $g_1(\mu) = \mathtt{sym}$ lorsqu'elle est symétrique.

Les mesures se répartissent en deux classes, présentées table III.2

$g_1(\mu) = \mathtt{sym}$	$g_1(\mu) = \mathtt{ori}$
R, GI, IQC, Lift, PS, Sup	ConfCen, Conf, Conv, -IndImp, IPD, IntImp, TIIE, Lap, Loe, FB, Seb, MoCo, Tec, Zhang

Tab. III.2 – Répartition des mesures pour les deux modalités de g_1

Dans d'autres contextes (comme dans le cadre de la validation d'arbres de décisions), d'autres propriétés portant sur l'évaluation différenciée d'une règle et d'une de ses formes complémentaires (par exemple $\mathtt{A} \rightarrow \mathtt{B}$ et $\mathtt{A} \rightarrow \bar{\mathtt{B}}$) peuvent faire sens.

Toutefois, dans le cadre des règles d'association, il est impossible de considérer les formes complémentaires aux itemsets, puisqu'on ne s'intéresse qu'aux co-présences. Ainsi, la "règle"

IRIN – Nantes) de l'Action Spécifique Stic 20 GafoDonnées (animée par Rosine Cicchetti, LIF – Marseille, et Michèle Sebag, LRI – Orsay) pour les discutions et les échanges fructueux autour du thème de la qualité. Nous espérons avoir également apporté notre contribution à ce groupe

$\mathtt{A} \rightarrow \bar{\mathtt{B}}$ ne peut être extraite. Nous écartons donc les considérations similaires à g_1 portant sur les formes complémentaires suivantes, sans toutefois ignorer qu'elles pourraient faire sens dans d'autres contextes :

- $\mathtt{B} \rightarrow \mathtt{A}$,
- $\mathtt{A} \rightarrow \bar{\mathtt{B}}$,
- $\bar{\mathtt{B}} \rightarrow \bar{\mathtt{A}}$.

Notons à nouveau que bien que nous écartions ces catégories de règles, cela n'empêche pas de considérer les cardinalités de ces complémentaires lorsque l'on souhaite évaluer la qualité d'une règle $\mathtt{A} \rightarrow \mathtt{B}$, puisque la donnée des fréquences p_a, p_b et p_{ab} suffit à reconstruire l'intégralité de la matrice de contingence de la table I.2. Ces compléments ont une sémantique utile, qu'il est important de considérer, tout particulièrement la notion de contre-exemples à une règle, $\mathtt{A}\bar{\mathtt{B}}$. Il convient donc de bien distinguer génération de règle et évaluation de règle.

2.2 Décroissance selon la taille du conséquent

Afin d'étudier la décroissance avec la taille du conséquent, nous raisonnons à nombre d'exemples et de contre-exemples fixés. Dans cette configuration, on peut considérer qu'une règle sera d'autant plus intéressante que n_b est faible [Piatetsky-Shapiro, 1991], et ainsi que la décroissance en fonction de la taille du conséquent peut constituer un critère de choix de mesure pertinent.

Les effectifs n_{ab} et $n_{a\bar{b}}$ du tableau de contingence I.2 étant fixés, n_a l'est également. Seuls les effectifs $n_{\bar{a}b}$ et $n_{\bar{a}\bar{b}}$ évoluent en fonction de n_b. Plus n_b est faible, plus $n_{\bar{a}b}$ l'est également : la partie de $\mathtt{B}$ non couverte par $\mathtt{A}$ se restreint. Il peut sembler logique de considérer ainsi, toutes choses étant égales par ailleurs, que plus la taille du conséquent d'une règle est faible, mieux la règle caractérise le conséquent, comme illustré en figure 2.2. Les mesures décroissantes avec n_b prendront cette notion en compte.

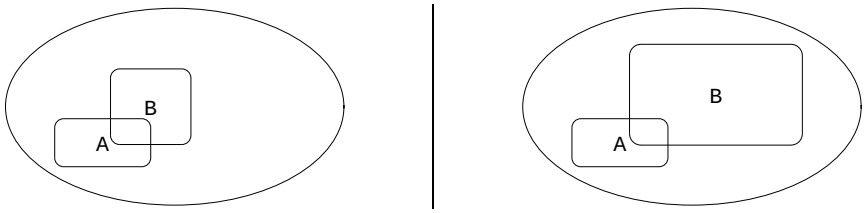


FIG. III.8 – Comparaison d'une règle à fort conséquent avec une règle à faible conséquent

Par ailleurs, si l'on s'intéresse à la découverte de "pépites", il est important qu'une mesure mette en avant les règles les moins générales (*i.e.* couvrant moins d'objets) dans les données. On retrouve en particulier cette logique de ciblage dans le cas de mesures centrant additivement ou multiplicativement la confiance sur p_b.

Nous considérons la propriété g_2, *décroissance avec* n_b, et lui attribuons deux modalités : $\mathtt{dec}_{p_b}$ lorsqu'une mesure met l'accent sur les règles à faible conséquent, et $\mathtt{non-dec}_{p_b}$ dans le cas contraire. La table III.3 présente la répartition des mesures selon ces deux modalités.

$g_2(\mu) = \texttt{dec}_{p_b}$	$g_2(\mu) = \texttt{non} - \texttt{dec}_{p_b}$
R, ConfCen, PS, Loe, Zhang, -IndImp, Lift, FB, Conv, GI, IQC, IntImp, TIIE, IPD	Sup, Conf, Lap, Seb, MoCo, Tec

Tab. III.3 – Répartition des mesures pour les deux modalités de g_2

2.3 Evaluation de situations de références

L'une des utilisations couramment faites des mesures de qualité, outre l'ordonnancement d'un ensemble de règles, est la sélection d'un sous-ensemble de règles. Face au trop grand nombre de règles extraites automatiquement, une stratégie usuelle permettant de limiter le nombre de règles qu'il sera nécessaire d'étudier consiste à les filtrer.

Ainsi, pour une mesure μ, l'expert des données fixe un seuil σ, et ne sont retenues pour validation manuelle que les règles r telles que $\mu(r) \geq \sigma$, les autres étant rejetées.

Dès lors, la problématique est donc de savoir comment placer correctement ce seuil σ. S'il est trop bas, il restera trop de règles après l'opération de filtrage pour que l'expert puisse procéder à une validation individuelle de chacune d'elles. S'il est fixé trop haut au contraire, on risque d'éliminer des règles intéressantes, voire même n'en retenir aucune !

Le fait de connaître *à priori* les valeurs prises par une mesure dans diverses situations de référence permet, lorsque ces valeurs sont constantes, de mieux appréhender le positionnement du seuil σ, dont la valeur reste généralement fixée en procédant par essai-erreur.

Nous en présentons quatre ci-dessous, dont trois retiendront notre attention dans notre contexte.

2.3.1 Règle logique

Il peut être intéressant de fixer un seuil de filtre en dessous de la valeur minimale que prend la mesure de qualité lorsqu'il n'y a pas de contre-exemples, afin d'éviter de rejeter ces règles. Les mesures étudiées étant des fonctions décroissantes de $n_{a\bar{b}}$, une valeur fixe (1 par exemple), ou à la rigueur infinie, peut ainsi être souhaitée dans le cas d'une règle logique. Bien que cette valeur fixe puisse poser problème (perte du pouvoir discriminant de la mesure), elle est fréquemment souhaitée [Lenca et al., 2003a]. La propriété g_3 prend deux valeurs : $g_3(\mu) = \texttt{max} - \texttt{var}$ si la mesure n'a pas une valeur fixe en cas de règle logique et $g_3(\mu) = \texttt{max} - \texttt{fixe}$ sinon.

La table III.4 donne la répartition des mesures selon ces deux modalités.

$g_3(\mu) = \texttt{max} - \texttt{fixe}$	$g_3(\mu) = \texttt{max} - \texttt{var}$
Conf, Conv, Loe, Seb, FB, Tec, Zhang, IntImp, TIIE	Sup, R, ConfCen, PS, GI, Lift, Lap, MoCo, IQC, -IndImp, IPD

Tab. III.4 – Répartition des mesures selon les deux modalités de g_3

2.3.2 Indépendance

Afin de ne conserver que les règles faisant sens d'un point de vue statistique, il peut être très intéressant de fixer un seuil de filtre au dessus de la valeur maximale que prend la mesure pour deux itemsets statistiquement indépendants.

Ceci se fait très aisément pour certaines mesures qui, comme nous l'avons vu au chapitre II s'annulent, ou tout du moins prennent une valeur fixe à l'indépendance (*i.e.* lorsque $p_{ab} = p_a p_b$). Nous considérons donc la propriété g_4 [Piatetsky-Shapiro, 1991], qui prend deux modalités : $g_4(\mu) = \texttt{indep} - \texttt{fixe}$ lorsque la mesure a une valeur constante à l'indépendance, et $g_4(\mu) = \texttt{indep} - \texttt{var}$ sinon. Le tableau III.5 donne la répartition des mesures selon les deux modalités de g_4.

$g_4(\mu) = \texttt{indep} - \texttt{fixe}$	$g_4(\mu) = \texttt{indep} - \texttt{var}$
R, ConfCen, Conv, PS, Loe, GI, Lift, FB, IQC, Zhang, -IndImp, IntImp, TIIE, IPD	Conf, Sup, Seb, Tec, Lap, MoCo

Tab. III.5 – Répartition des mesures selon les deux modalités de g_4

La table III.6 croise les valeurs de g_3 et g_4 pour nos 20 mesures. Seules 5 mesures ont à la fois une valeur fixe à l'indépendance, et pour la règle logique. Toutefois, plusieurs mesures récentes, telle IPEE, viennent augmenter le nombre de mesures étant fixes pour ces deux situations de références.

	$g_4(\mu) = \texttt{indep} - \texttt{fixe}$	$g_4(\mu) = \texttt{indep} - \texttt{var}$
$g_3(\mu) = \texttt{max} - \texttt{fixe}$	Conv, Loe, FB, Zhang, IntImp	Conf, Seb, Tec, TIIE
$g_3(\mu) = \texttt{max} - \texttt{var}$	R, ConfCen, PS, GI, Lift, IQC, -IndImp, IPD	Sup, Lap, MoCo

Tab. III.6 – Répartition des mesures selon g_3 et g_4

2.3.3 Indétermination

Une autre valeur de référence récemment mise en avant est celle qui intervient lorsqu'une règle a autant d'exemples que de contre-exemples. Dans ce cas, il y a indétermination [Blanchard et al., 2005a]. De même qu'à l'indépendance, on peut souhaiter ne pas retenir les règles au-delà de cette limite, et donc fixer un seuil adapté. Ceci se fera plus facilement si la mesure prend une valeur fixe à cette référence. On note g_5 cette propriété, qui prend deux modalités : $g_5(\mu) = \texttt{indet} - \texttt{fixe}$ lorsque la mesure prend une valeur constante à l'indétermination, et $g_5(\mu) = \texttt{indet} - \texttt{var}$ sinon.

Le tableau III.7 donne la répartition des mesures selon les deux modalités de g_5.

Notons que de toutes nos mesures, seule TIIE possède une valeur fixe pour les trois situations de référence.

$g_5(\mu) = \texttt{indet} - \texttt{fixe}$	$g_5(\mu) = \texttt{indet} - \texttt{var}$
Conf, Seb, MoCo, Tec, TIIE	Sup, r, ConfCen, Conv, PS, Loe, GI, Lift, Lap, FB, IQC, Zhang, -IndImp, IntImp, IPD

Tab. III.7 – Répartition des mesures selon les deux modalités de g_5

2.3.4 Incompatibilité

Nous écartons de notre étude la propriété constituée par l'évaluation de l'incompatibilité. Cette quatrième situation de référence définie par $\texttt{A} \cap \texttt{B} = \emptyset$, qui exprime que $\texttt{B}$ ne peut pas se réaliser si $\texttt{A}$ est réalisé, n'intéresse pas le cadre des règles d'association. En effet, cette situation intéresse la règle $\texttt{A} \rightarrow \overline{\texttt{B}}$ qui est alors une règle logique et non pas la règle $\texttt{A} \rightarrow \texttt{B}$.

A nouveau, dans d'autres paradigmes de fouille de données, ou dans un contexte d'extraction de règles d'association négatives, cette propriété pourrait tout à fait faire sens.

2.4 Tolérance aux premiers contre-exemples

La tolérance aux premiers contre-exemples, c'est-à-dire la décroissance rapide ou non d'une mesure autour de 0^+ en fonction de $n_{a\bar{b}}$, est une propriété qui peut être souhaitée ou non, en fonction des besoins de $E_{\mathcal{A}}$ [Gras et al., 2001].

Ce sera le cas dans des applications de marketing, par exemple : il n'est pas rare de voire que parmi une population ciblée par une offre promotionnelle, nombreux sont les individus qui ne sont pas intéressés. Si le coût financier d'une campagne peut être réduit en améliorant le taux de réponse, il n'en reste pas moins certain que commettre quelques erreurs ne sera pas trop préjudiciable. A l'opposé, si l'on considère une application au domaine médical, il peut tout à fait être judicieux de souhaiter qu'une règle soit fortement pénalisée dès que quelques individus sont en contradiction avec ce qu'elle stipule, afin de minimiser les risques thérapeutiques.

Nous avons retenu trois modalités pour ce critère (linéarité, concavité ou convexité), et étudions dans un premier temps nos mesures selon la modélisation classiquement utilisée. D'autres modélisations sont toutefois envisageables, et nous proposons dans un second temps les résultats qui y sont liés pour un nombre restreint de mesures.

2.4.1 Modélisation classique

La propriété g_6 s'intéresse ainsi au comportement des mesures de qualité en fonction des premiers contre-exemples, à marges fixées (*i.e.* p_a et p_b étant constants), et nous proposons trois modalités :

- tolérance à l'apparition des premiers contre-exemples (une telle mesure de qualité sera donc concave en fonction de $p_{a\bar{b}}$ au voisinage de 0^+),
- rejet dès les premiers contre-exemples (une telle mesure de qualité sera donc convexe en fonction de $p_{a\bar{b}}$ au voisinage de 0^+),

- indifférente (une telle mesure de qualité sera linéaire en fonction de $p_{a\bar{b}}$ au voisinage de 0^+)

Il faut remarquer que dans la mesure où la confiance est une transformée affine du nombre d'exemples ou de contre-exemples à marges fixées, toutes les mesures qui sont des transformées affines de la confiance ont cette même caractéristique.

On notera $g_6(\mu) = \texttt{conv}$ si μ est convexe pour $n_{a\bar{b}}$ voisin de 0^+, $g_6(\mu) = \texttt{lin}$ si elle est linéaire et $g_6(\mu) = \texttt{conc}$ si elle est concave. La table III.8 classe les mesures selon les trois modalités de g_6. Ces variations s'évaluent en raisonnant à marges fixées.

$g_6(\mu) = \texttt{conv}$	$g_6(\mu) = \texttt{lin}$	$g_6(\mu) = \texttt{conc}$
FB, Conv, Seb	ConfCen, Conf, r, -IndImp, IQC, Lap, Lift, Loe, PS, Sup, MoCo	GI, IPD, IntImp, TIIE, Tec, Zhang

Tab. III.8 – Répartition des mesures selon les trois modalités de g_6

2.4.2 Autres modélisations envisageables

Dans [Vaillant et al., 2006a], nous proposons d'étudier d'autres modélisations de l'introduction de contre-exemples. Dans la modélisation classique, on travaille à marge fixées : p_a et p_b sont constants et les cellules internes de la matrice croisant `A` et `B` varient. Toutefois, cette introduction de contre-exemples basée sur une translation des ensembles `A` et `B` (voire figure III.9 et tableau III.9) n'est pas nécessairement la seule qui peut se justifier. Si l'on sait prédire d'où proviennent les contre-exemples, le recours à une autre modélisation peut se justifier, et les résultats de l'évaluation de la propriété g_6 en sont affectés.

Par exemple, si l'on étudie le rayonnement cosmique de corps célestes, on sait que la proximité d'une masse provoque ce que l'on appelle un "décalage gravitationnel vers le rouge". Ce phénomène, que l'on peut quantifier en théorie de la relativité, a pour effet de décaler le spectre observé lorsque celui-ci passe à proximité d'astres. Si le décalage associé au Soleil est minime, de l'ordre de 1/100Å dans le spectre du visible, l'impact d'autres corps est bien plus important. Par exemple, l'étoile naine blanche Sirius B, compagnon de Sirius, entraîne un décalage d'environ 1Å à la longueur d'onde 5000Å. On sait ainsi que ce phénomène entraînera un décalage orienté des observations, et on peut éventuellement en déduire une typologie des erreurs liées. Rien ne laisse à penser que dans ce contexte, l'erreur introduite se fait à marges fixées.

Ainsi, nous introduisons deux modélisations alternatives à la modélisation classique, et observons l'impact du choix de la modélisation. De manière simple, nous augmentons la proportion de contre-exemples $p_{a\bar{b}}$ (initialement nulle) d'une quantité x, et observons les variations de plusieurs mesures en fonction de x.

Dans la modélisation 2, nous considérons que l'ajout de contre-exemples se fait en diminuant p_{ab}. Cette fois, les marges ne sont plus fixées. La figure III.10 et la table III.10 présentent l'évolution d'une règle $\texttt{A} \rightarrow \texttt{B}$ selon cette seconde stratégie.

Enfin, la modélisation 3 se fait elle en intégrant de nouveaux contre-exemples à partir

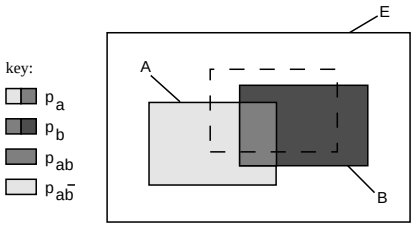


FIG. III.9 – Modélisation 1

A\B	0	1	total
0	$p_{\bar{b}} - x$	$p_{\bar{a}} - p_{\bar{b}} + x$	$p_{\bar{a}}$
1	x	$p_a - x$	p_a
total	$p_{\bar{b}}$	p_b	1

TAB. III.9 – Table de contingence associée à la modélisation 1

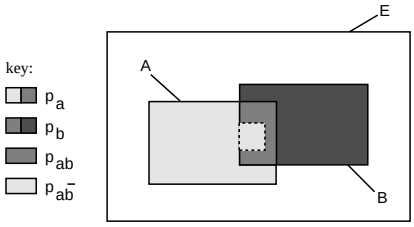


FIG. III.10 – Modélisation 2

A\B	0	1	total
0	$p_{\bar{a}\bar{b}}$	$p_{\bar{a}b}$	$p_{\bar{a}}$
1	x	$p_a - x$	p_a
total	$p_{\bar{a}\bar{b}} + x$	$p_{\bar{a}b} + p_a - x$	1

TAB. III.10 – Table de contingence associée à la modélisation 2

des objets non-couverts par la règle. La figure III.11 et la table III.11 présentent l'évolution d'une règle $\mathtt{A} \rightarrow \mathtt{B}$ selon cette dernière stratégie.

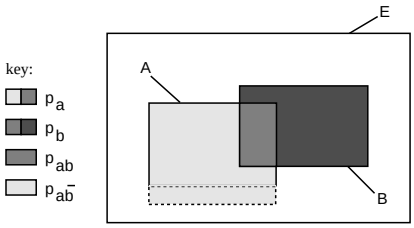


FIG. III.11 – Modélisation 3

A\B	0	1	total
0	$p_{\bar{b}} - x$	$p_{\bar{a}b}$	$p_{\bar{b}} + p_{\bar{a}b} - x$
1	x	p_{ab}	$p_{ab} + x$
total	$p_{\bar{b}}$	p_b	1

TAB. III.11 – Table de contingence associée à la modélisation 3

D'autres modélisations ont été envisagées, mais n'ont pas été explorées car semblant moins pertinentes. Ainsi, nous aurions également pu considérer :

- que les nouveaux contre-exemples puissent provenir de $\bar{\mathtt{A}}\mathtt{B}$. Dans ce cas, les grandeurs p_a et p_b sont toutes deux variables, l'ensemble $\mathtt{A}$ augmentant à mesure que $\mathtt{B}$ diminue.
- que les nouveaux contre-exemples proviennent de nouveaux cas. C'est alors n qui évolue, à mesure que l'on ajoute des exceptions à la règle.

L'étude des dérivées premières et secondes en fonction de x présentée dans [Vaillant et al., 2006a] permet de conclure, pour un nombre restreint de mesures, d'une part que quelque soit la modélisation choisie, ces mesures sont décroissantes du nombre de contre-exemples. D'autre part, que la modélisation retenue a un impact sur les variations des mesures en fonction de $p_{a\bar{b}}$ autour de 0^+. La table III.12 synthétise les résultats obtenus pour les trois modélisations.

	Modélisation 1	**Modélisation 2**	**Modélisation 3**
CONF	linéaire	linéaire	convexe
CONV	convexe	convexe	convexe
TEC	concave	concave	linéaire
JAC	convexe	linéaire	convexe
MOCO	linéaire	concave	linéaire
LIFT	linéaire	concave	convexe
GI	concave	concave	convexe
LOE	linéaire	convexe	convexe
PS	linéaire	linéaire	linéaire
SUP	linéaire	linéaire	non affecté

TAB. III.12 – Impact du choix de la modélisation sur l'évaluation de la propriété g_6

Dans la suite de ce mémoire, nous utiliserons la modélisation 1 pour évaluer la propriété g_6, car cette modélisation nous semble la plus juste en l'absence de connaissance sur l'origine des contre-exemples. Cette évaluation est toutefois à réviser si l'on est en mesure de modéliser la provenance de ces contre-exemples. Seule PS est insensible au choix de la modélisation.

2.5 Prise en compte du nombre de cas

Intuitivement, pour des proportions relatives identiques dans le tableau de contingence de $\mathtt{A} \times \mathtt{B}$, la règle est d'autant plus fiable que le nombre de transactions n est élevé. On peut légitimement souhaiter que la mesure soit croissante lorsque l'on dilate les données en augmentant n, toutes proportions fixées [Gras, 1979, Lallich, 2002, Gras et al., 2004].

L'inconvénient de la prise en compte de n dans une telle mesure est que lorsque n croît, celle-ci évalue de mieux en mieux un grand nombre de règles. Ainsi, ces évaluations peuvent s'approcher du maximum, et les mesures bornées supérieurement perdent leur pouvoir discriminant. C'est ainsi qu'une mesure purement descriptive pour laquelle n intervient dans le calcul du seuil à partir duquel on retient la règle peut être préférée. Le fait de satisfaire à la prise en compte du nombre de cas s'accompagne d'une perte de capacité discriminante.

On note `desc` si la mesure est invariante en cas de dilatation des données et `stat` si elle est croissante en cas de dilatation des données.

$g_7(\mu) = \mathtt{desc}$	$g_7(\mu) = \mathtt{stat}$
Sup, Conf, ConfCen, Conv, PS, Loe, GI, Seb, Lift, MoCo, FB, Tec, IQC, Zhang, Lap, r, -IndImp	IntImp, TIIE, IPD

Tab. III.13 – Répartition des mesures selon les deux modalités de g_7

2.6 Facilité à fixer un seuil d'acceptation de la règle

La facilité à placer un seuil d'acceptation, par exemple par rapport à une situation de référence [Lenca et al., 2003b], est aussi selon nous une propriété importante. En effet, lorsque l'utilisateur cherche à fixer un seuil d'acceptation, peut-il estimer la fiabilité de ce seuil ? Ainsi, si l'on utilise Lift afin de filtrer un ensemble de règles, un seuil minimal de 1 permet de garantir que l'on ne retiendra que les règles corrélées positivement. Mais le seront-elles suffisamment pour qu'il soit pertinent de les retenir ? Si l'on augmente ce seuil à 1.2, ne risque-t-on alors pas de rejeter des règles intéressantes ?

Nous étudions dans [Lenca et al., 2006] l'impact qu'a le retrait d'exemples à une règle, étant donné une mesure et un seuil d'acceptation. Au-delà de cette étude simplifiée, [Lallich and Teytaud, 2004] proposent d'utiliser un paramètre de contrôle afin de résoudre ce problème de fixation de seuil.

2.6.1 Premières considérations

Nous considérons pour cette étude une règle $\mathtt{A} \rightarrow \mathtt{B}$, évaluée par une mesure μ au-delà d'un seuil σ_μ. Cette règle sera donc acceptée si l'on procède à une opération de filtrage, *i.e.* on suppose que $\mu(\mathtt{A} \rightarrow \mathtt{B}) \geq \sigma_\mu$. On souhaite estimer la proportion de contre-exemples qu'il est nécessaire d'ajouter à la règle pour que son évaluation ne soit plus considérée comme suffisante pour la retenir.

Pour illustrer la problématique, nous nous donnons deux règles $\mathtt{r_1}$ et $\mathtt{r_2}$, dont les caractéristiques sont données en tables III.14 et III.15. Les évaluations de ces règles par SUP, CONF et LIFT sont données en table III.16. Clairement, les valeurs de SUP et CONF sont suffisantes pour que de telles règles puissent être extraites via un algorithme de type APRIORI. De plus, elles ont toutes deux un LIFT supérieur à 1, et seraient ainsi retenues si l'on se donne cette valeur comme seuil de filtrage. Pour autant, $\mathtt{r_1}$, qui a une CONF plus faible, peut perdre 25% de ses exemples, et son évaluation par LIFT restera supérieure à 1, alors que $\mathtt{r_2}$ ne peux perdre que 20% d'exemples. De ce point de vue, $\mathtt{r_1}$ est plus robuste que $\mathtt{r_2}$, lorsqu'elles sont évaluées en post-analyse par LIFT, avec un seuil de 1. La robustesse des règles finalement retenues par une mesure, étant donné un seuil, a été peu étudiée. Nous proposons quelques éléments de réponses pour des transformées de CONF à marges fixées. Ce critère reste encore à formaliser et explorer [Lenca et al., 2006].

A\B	0	1	total
0	2000	2000	4000
1	2000	4000	6000
total	4000	6000	10000

TAB. III.14 – Caractéristiques de $\mathtt{r_1}$

A\B	0	1	total
0	1800	3200	5000
1	1000	4000	5000
total	2800	7200	10000

TAB. III.15 – Caractéristiques de $\mathtt{r_2}$

	SUP	CONF	LIFT
$\mathtt{r_1}$	0.40	0.66	1.11
$\mathtt{r_2}$	0.40	0.80	1.11

TAB. III.16 – Evaluation des règles $\mathtt{r_1}$ et $\mathtt{r_2}$

L'étude des transformées de la confiance met en évidence la proportion d'exemples qu'une règle peut perdre tout en restant valide.

A partir d'une règle $\mathtt{A} \rightarrow \mathtt{B}$, nous construisons la règle $\mathtt{A'} \rightarrow \mathtt{B'}$, correspondant à la règle $\mathtt{A} \rightarrow \mathtt{B}$ à laquelle on retranche une proportion α d'exemples, que nous transférons parmi les contre-exemples. La table III.17 croise les fréquences des ensembles $\mathtt{A'}$ et $\mathtt{B'}$ ainsi construits.

La figure III.12 illustre cette transformation. On note μ_σ^+ l'ensemble des règles vérifiant $\mu(\mathtt{A} \rightarrow \mathtt{B}) \geq \sigma$.

$\mathtt{A}'\backslash\mathtt{B}'$	0	1	total
0	$p_{\bar{a}\bar{b}}$	$p_{\bar{a}b}$	$p_{\bar{a}}$
1	$p_{a\bar{b}} + \alpha p_{ab}$	$p_{ab} - \alpha p_{ab}$	p_a
total	$p_{\bar{b}} + \alpha p_{ab}$	$p_b - \alpha p_{ab}$	1

TAB. III.17 – Caractéristiques de $\mathtt{A}' \rightarrow \mathtt{B}'$

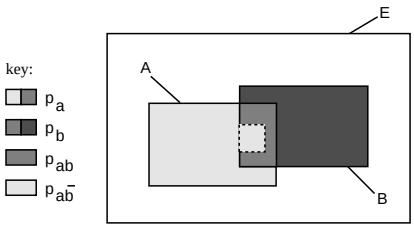


FIG. III.12 – Caractéristiques de $\mathtt{A}' \rightarrow \mathtt{B}'$

Dans le cas où la mesure peut s'écrire $\mu = \theta_1(\text{CONF} - \theta_0)$, avec θ_0 et θ_1 ne dépendant que des fréquences marginales, on montre dans [Lenca et al., 2006] qu'alors la proportion α d'exemples que l'on peut retrancher vérifie :

$$\mathtt{A}' \rightarrow \mathtt{B}' \in \mu_\sigma^+ \Leftrightarrow \alpha \leq \frac{(\mu - \sigma_\mu) + \sigma_\mu(1 - \frac{\theta_1}{\theta_1'})}{\mu + \theta_1\theta_0} + \theta_1 \frac{\theta_0 - \theta_0'}{\mu + \theta_1\theta_0}$$

où μ est la valeur que prend la mesure pour la règle $\mathtt{A} \rightarrow \mathtt{B}$, et θ_0' et θ_1' les coefficients linéaires associés à la règle $\mathtt{A}' \rightarrow \mathtt{B}'$.

La table III.18 (respectivement III.19) donne l'expression des coefficients linéaires associés à $\mathtt{A} \rightarrow \mathtt{B}$ (respectivement $\mathtt{A}' \rightarrow \mathtt{B}'$). La table III.20 donne la valeur maximale de α associée à ces cas de figure, pour neuf mesures pouvant s'écrire selon cette décomposition.

On montre également que si μ s'écrit $f(\text{CONF})$, f étant une fonction monotone croissante inversible, ne faisant intervenir que les fréquences marginales dans sa définition, alors :

$$\mathtt{A}' \rightarrow \mathtt{B}' \in \mu_\sigma^+ \Leftrightarrow \alpha \leq \frac{\text{CONF} - f^{-1}(\sigma_\mu)}{\text{CONF}}$$

où f^{-1} est la fonction réciproque de f.

La table III.21 présente quatre mesures se prêtant à cette écriture, et la table III.22 les valeurs maximales de α associées.

A partir de ces résultats, on peut étudier la sensibilité du paramètre α en fonction de la règle $\mathtt{A} \rightarrow \mathtt{B}$. Le tableau III.23 présente les variations de $\theta_{\max}$ en fonction de p_a, p_b et p_{ab}. Ce dernier est défini comme le coefficient tel que $\alpha_{\max} = \theta_{\max}(\mu - \sigma)$, et correspond donc aux variations de $\alpha_{\max}$ pour un écart fixé entre la valeur prise par la mesure et le seuil considéré.

	Définitions relatives	θ_0	θ_1
ConfCen	$p_{b/a} - p_b$	p_b	1
Conf	$p_{b/a}$	0	1
Gan	$2p_{b/a} - 1$	0.5	2
Lap	$\frac{p_{b/a}+\frac{1}{np_a}}{1+\frac{2}{np_a}}$	$-\frac{1}{np_a}$	$\frac{1}{1+\frac{2}{np_a}}$
MoCo	$2\frac{p_a}{p_b}(p_{b/a} - 0.5)$	0.5	$2\frac{p_a}{p_b}$
Lift	$\frac{p_{b/a}}{p_b}$	0	$\frac{1}{p_b}$
Loe	$\frac{p_{b/a}-p_b}{1-p_b}$	p_b	$\frac{1}{1-p_b}$
PS	$n(p_{ab} - p_a p_b)$	p_b	np_a
Sup	p_{ab}	0	p_a

Tab. III.18 – Transformations linéaires de la confiance

μ	θ'_0	θ'_1
ConfCen	$p_b - \alpha p_{ab}$	$\theta'_1 = \theta_1 = 1$
Gan	$\theta'_0 = \theta_0 = 0.5$	$\theta'_1 = \theta_1 = 2$
Lap	$\theta'_0 = \theta_0 = -\frac{1}{np_a}$	$\theta'_1 = \theta_1 = \frac{1}{1+\frac{2}{np_a}}$
MoCo	$\theta'_0 = \theta_0 = 0.5$	$2\frac{p_a}{p_b-\alpha p_{ab}}$
Lift	$\theta'_0 = \theta_0 = 0$	$\frac{1}{p_b-\alpha p_{ab}}$
Loe	$p_b - \alpha p_{ab}$	$\frac{1}{1-p_b+\alpha p_{ab}}$
PS	$p_b - \alpha p_{ab}$	$\theta'_1 = \theta_1 = np_a$
Sup	$\theta'_0 = \theta_0 = 0$	$\theta'_1 = \theta_1 = p_a$

Tab. III.19 – Coefficients θ'_0 et θ'_1

μ	α
ConfCen	$\frac{\text{ConfCen}-\sigma}{\text{ConfCen}+p_{\bar{a}b}}$
Conf	$\frac{\text{Conf}-\sigma}{\text{Conf}}$
Gan	$\frac{\text{Gan}-\sigma}{\text{Gan}+1}$
Lap	$\frac{\text{Lap}-\sigma}{\text{Lap}-\frac{1}{2+np_a}}$
MoCo	$\frac{\text{MoCo}p_b-\sigma p_b}{\text{MoCo}p_b+p_a-\sigma p_{ab}}$
Lift	$\frac{\text{Lift}-\sigma}{\text{Lift}(1-\sigma p_a)}$
Loe	$\frac{(\text{Loe}-\sigma)(1-p_b)}{\text{Loe}(1-p_b)+p_b-(1-\sigma)p_{ab}}$
PS	$\frac{\text{PS}-\sigma}{np_{ab}(1-p_a)}$
Sup	$\frac{\text{Sup}-\sigma}{\text{Sup}}$

Tab. III.20 – Valeur maximale de α, cas des transformées linéaires

	Définitions	$f(\text{Conf})$
Conv	$\frac{p_a p_{\bar{b}}}{p_{a\bar{b}}}$	$\frac{p_{\bar{b}}}{1-\text{Conf}}$
Tec	$1-\frac{p_{a\bar{b}}}{p_{ab}}$	$\frac{2\text{Conf}-1}{\text{Conf}}$
Jac	$\frac{p_{ab}}{p_a+p_b-p_{ab}}$	$\frac{\text{Conf}}{1+\frac{p_b}{p_a}-\text{Conf}}$
Seb	$\frac{p_{ab}}{p_{a\bar{b}}}$	$\frac{\text{Conf}}{1-\text{Conf}}$

Tab. III.21 – Transformées monotones de Conf

μ	α
Conv	$\frac{p_a-p_{ab}}{p_{ab}}\frac{\text{Conv}-\sigma}{\sigma-p_a}$
Tec	$\frac{\text{Tec}-\sigma}{2-\sigma}$
Jac	$1-\frac{\sigma}{\text{Jac}}$
Seb	$\frac{\text{Seb}-\sigma}{\text{Seb}(1+\sigma)}$

Tab. III.22 – Valeurs maximales de α, cas des transformées monotones

A partir de ces résultats, on peut alors connaître la robustesse d'une règle, pour ces treize mesures, sans coût de calcul supplémentaire sur les données, puisque la valeur maximale de α est exprimée en fonction de grandeurs déjà connues. Cette valeur permet ainsi de quantifier numériquement la résistance qu'aurait une règle à l'introduction de bruit. La résistance d'une règle au bruit a particulièrement été étudiée par [Azé, 2003].

	$\theta_{\max}$	p_a	p_b	p_{ab}
ConfCen	$\frac{p_a}{p_{ab}-p_a p_{ab}}$	$\nearrow$	$=$	$\searrow$
Conf	$\frac{p_a}{p_{ab}}$	$\nearrow$	$=$	$\searrow$
Gan	$\frac{p_a}{2p_{ab}}$	$\nearrow$	$=$	$\searrow$
Lap	$\frac{2+np_a}{np_{ab}}$	$\nearrow$	$=$	$\searrow$
MoCo	$\frac{p_b}{p_{ab}\frac{p_a+p_b}{p_{ab}}-1}$	$=$	$\nearrow$	$\searrow$
Lift	$\frac{p_a p_b}{p_{ab}(1-\sigma p_a)}$	$\nearrow$	$\nearrow$	$\searrow$
Loe	$\frac{p_a(1-p_b)}{p_{ab}(1-p_a(1-\sigma))}$	$\nearrow$	$\searrow$	$\searrow$
PS	$\frac{1}{np_{ab}(1-p_a)}$	$\nearrow$	$=$	$\searrow$
Sup	$\frac{1}{p_{ab}}$	$=$	$=$	$\searrow$
Conv	$\frac{p_a-p_{ab}}{p_{ab}(\sigma-p_a)}$	$\nearrow$	$=$	$\searrow$
Tec	$\frac{1}{2-\sigma}$	$=$	$=$	$=$
Jac	$\frac{p_a+p_b}{p_{ab}}-1$	$\nearrow$	$\nearrow$	$\searrow$
Seb	$\frac{p_a}{p_{ab}(1+\sigma)}-\frac{1}{1+\sigma}$	$\nearrow$	$=$	$\searrow$

Tab. III.23 – Variations de $\theta_{\max}$ en fonction de p_a, p_b et p_{ab}

2.6.2 Utilisation d'un paramètre de contrôle

Face à la multitude de règles évaluées, il est important de pouvoir facilement fixer le seuil à partir duquel on considère que les règles ont un réel intérêt, sans avoir à les classer.

Une autre manière classique de procéder consiste à se référer à la probabilité critique de la valeur observée de la mesure sous l'hypothèse d'indépendance (ou *p-value*). Celle-ci ne doit pas être interprétée comme un risque statistique, compte tenu de la multitude de tests effectués [Lallich and Teytaud, 2004], mais comme un paramètre de contrôle. La fixation d'un seuil est alors immédiate lorsque la mesure est elle-même définie comme une telle probabilité critique. C'est le cas de IntImp, ou IPD. Sinon, on peut le plus souvent définir un tel seuil en partant de l'un ou l'autre des trois types de modélisation proposés par [Lerman et al., 1981] pour établir la loi de n_{ab} sous l'hypothèse d'absence de lien. Les seules mesures qui se prêtent mal à un tel calcul sont Zhang, en raison de son dénominateur et TIIE en raison de la pondération par l'entropie de l'expérience. On note $g_8(\mu) = \texttt{non} - \texttt{pc}$ si la mesure se prête mal au calcul d'un seuil (dans le cas de Zhang et TIIE) et $g_8(\mu) = \texttt{pc}$ si elle se prête bien à un tel calcul (les 18 autres mesures).

2.7 Intelligibilité de la mesure

Dans certaines situations, notamment pour communiquer et expliquer des résultats, l'intelligibilité de la mesure a de l'importance [Lenca et al., 2003a]. On considérera que la mesure est intelligible si son sens concret peut s'écrire par une phrase simple. On note $g_9(\mu) = a$ s'il y a impossibilité de donner le sens concret de la mesure, $g_9(\mu) = b$ si la mesure se ramène à des quantités usuelles et $g_9(\mu) = c$ si le sens concret de la mesure s'écrit en une phrase.

Le tableau III.24 liste la répartition des diverses mesures se répartissent selon les trois modalités de g_9.

$g_9(\mu) = a$	$g_9(\mu) = b$	$g_9(\mu) = c$
Zhang, -IndImp, IQC, GI, IntImp, TIIE, Lap, IPD	r, PS, Loe, MoCo, Seb, Tec, Conv	FB, Sup, Conf, ConfCen Lift

Tab. III.24 – Répartition des mesures selon les trois modalités de g_9

2.8 Analyse de l'espace des mesures

Le tableau III.25 résume les différentes propriétés utilisées dans notre étude. Si on adjoint à ces propriétés un ordre sur les modalités qu'elles prennent, nous obtenons une famille de critères, exploitable en Amd. Selon les propriétés, nous pensons qu'il faut alternativement prendre en compte des choix d'$E_{\mathcal{A}}$ (expert de l'analyse), lorsque ces propriétés semblent normatives, ou d'$E_{\mathcal{R}}$ (expert des règles), lorsque l'ordre que l'on peut souhaiter donner sur les modalités dépend bien du contexte applicatif. Cette séparation en deux catégories ne représente que le fait que l'ordre dans lequel on pourra ranger les différentes modalités des propriétés peut ne pas être totalement libre (dans le cas de propriétés normatives).

Notons toutefois qu'un critère peut ne pas être pris en compte dans une approche AMD si sa considération ne fait pas sens.

L'analyse AMD se fera donc en deux temps. Dans un premier temps, on ordonne les modalités des quatre propriétés non-normatives en suivant les préférences exprimées par $E_{\mathcal{R}}$. Dans un second temps, la détermination du jeu de poids associé aux neuf critères est discutée. Dans les deux cas, un échange entre $E_{\mathcal{A}}$ et $E_{\mathcal{R}}$ est nécessaire. Par exemple, il est particulièrement important que les poids traduisent bien l'importance relative de chaque critère telle que vue par $E_{\mathcal{R}}$.

La propriété g_8 revêt un caractère particulier, car elle dépend fortement du niveau d'expertise en statistiques d'$E_{\mathcal{R}}$. Ainsi, nous proposons un ordonnancement normatif de bon sens, tout en laissant l'option de remettre en cause cet ordre. Les trois autres critères subjectifs que nous proposons ($g6$, g_7 et g_9) sont beaucoup plus accessibles à $E_{\mathcal{R}}$.

Critère	Sémantique	Nombre de modalités	Responsabilité
g_1	non symétrie selon `A` et `B`	2	$E_{\mathcal{A}}$
g_2	décroissance avec n_b	2	$E_{\mathcal{A}}$
g_3	situation à la règle logique	2	$E_{\mathcal{A}}$
g_4	situation à l'indépendance	2	$E_{\mathcal{A}}$
g_5	situation à l'indétermination	2	$E_{\mathcal{A}}$
g_6	tolérance aux premiers contre-exemples	3	$E_{\mathcal{R}}$
g_7	prise en compte de n	2	$E_{\mathcal{R}}$
g_8	facilité à fixer un seuil	2	$E_{\mathcal{A}}$ et $E_{\mathcal{R}}$
g_9	intelligibilité	3	$E_{\mathcal{R}}$

TAB. III.25 – Famille $\mathcal{F}$ de critères

Nous introduisons les ordres suivants sur les propriétés normatives :

g_1 : $\texttt{ori} \succ \texttt{sym}$

g_2 : $\texttt{dec}_{p_b} \succ \texttt{non} - \texttt{dec}_{p_b}$

g_3 : $\texttt{max} - \texttt{fixe} \succ \texttt{max} - \texttt{var}$

g_4 : $\texttt{indep} - \texttt{fixe} \succ \texttt{indep} - \texttt{var}$

g_5 : $\texttt{indet} - \texttt{fixe} \succ \texttt{indet} - \texttt{var}$

g_8 : $\texttt{pc} \succ \texttt{non} - \texttt{pc}$

Nous rappelons de manière synthétique les évaluations des vingt mesures sur les neuf propriétés en table III.26. En encodant sous forme normale disjonctive cette matrice, on obtient la matrice booléenne III.27. On peut alors construire la classification ascendante hiérarchique (CAH) associée à cette matrice, et ainsi mieux mettre en évidence les groupes de mesures proches, en termes d'analyse formelle. Ces rapprochements peuvent être retrouvés dans le plan GAIA. Nous mettrons en regard ces regroupements de mesures selon l'approche formelle avec une classification obtenue via des comparaisons entre mesures basées sur des résultats expérimentaux au chapitre IV [Vaillant et al., 2004, Vaillant et al., 2006b]. La figure III.13 présente le dendrogramme obtenu sur la matrice III.27, en appliquant le critère de Ward, et en utilisant la distance de Manhattan. D'autres critères produisent des résultats semblables.

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
Sup	sym	non − dec$_{p_b}$	max − var	indep − var	indet − var	lin	desc	pc	c
Conf	ori	non − dec$_{p_b}$	max − fixe	indep − var	indet − fixe	lin	desc	pc	c
r	sym	dec$_{p_b}$	max − var	indep − fixe	indet − var	lin	desc	pc	b
ConfCen	ori	dec$_{p_b}$	max − var	indep − fixe	indet − var	lin	desc	pc	c
PS	sym	dec$_{p_b}$	max − var	indep − fixe	indet − var	lin	desc	pc	b
Loe	ori	dec$_{p_b}$	max − fixe	indep − fixe	indet − var	lin	desc	pc	b
Zhang	ori	dec$_{p_b}$	max − fixe	indep − fixe	indet − var	conc	desc	non − pc	a
- IndImp	ori	dec$_{p_b}$	max − var	indep − fixe	indet − var	lin	desc	pc	a
Lift	sym	dec$_{p_b}$	max − var	indep − fixe	indet − var	lin	desc	pc	c
MoCo	ori	non − dec$_{p_b}$	max − var	indep − var	indet − fixe	lin	desc	pc	b
Seb	ori	non − dec$_{p_b}$	max − fixe	indep − var	indet − fixe	conv	desc	pc	b
FB	ori	dec$_{p_b}$	max − fixe	indep − fixe	indet − var	conv	desc	pc	c
Conv	ori	dec$_{p_b}$	max − fixe	indep − fixe	indet − var	conv	desc	pc	b
Tec	ori	non − dec$_{p_b}$	max − fixe	indep − var	indet − fixe	conc	desc	pc	b
IQC	sym	dec$_{p_b}$	max − var	indep − fixe	indet − var	lin	desc	pc	a
GI	sym	dec$_{p_b}$	max − var	indep − fixe	indet − var	conc	desc	pc	a
IntImp	ori	dec$_{p_b}$	max − fixe	indep − fixe	indet − var	conc	stat	pc	a
TIIE	ori	dec$_{p_b}$	max − fixe	indep − fixe	indet − fixe	conc	stat	non − pc	a
IPD	ori	dec$_{p_b}$	max − var	indep − fixe	indet − var	conc	stat	pc	a
Lap	ori	non − dec$_{p_b}$	max − var	indep − var	indet − var	lin	desc	pc	a

Tab. III.26 – Matrice de synthèse sur les propriétés

Nous pouvons en dégager cinq classes principales :
- Sup, MoCo, Lap ;
- Conf, Seb, Tec ;
- r, PS, ConfCen, Lift, -IndImp, IQC, GI ;
- Loe, FB, Conv ;
- Zhang, TIIE, IntImp, IPD.

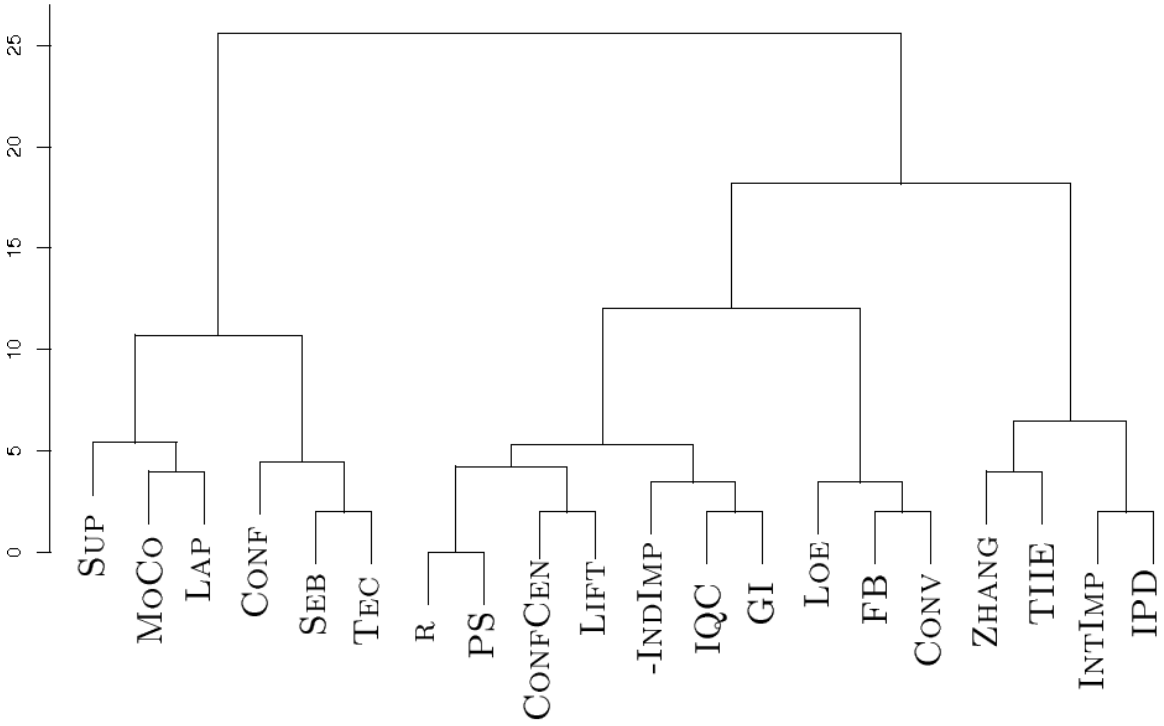


Fig. III.13 – Dendrogramme associé à la CAH des mesures selon leurs propriétés formelles

2.9 Mise en œuvre : deux scenarii orientés utilisateur

Nous allons comparer deux scenarii simples, Sc1 et Sc2. Nous proposons d'ordonner les critères selon des ordres de "bon" sens, et d'observer l'impact du choix d'un ordre différent sur l'un d'entre eux, g_6. Nous choisirons des jeux de poids égaux pour tous les critères, soit de 1/9 de manière normalisée.

Les deux scenarii que nous envisageons représentent deux choix classiques. L'expert des données peut en effet tolérer l'apparition de contre-exemples ou non, selon le caractère critique de l'existence de ces contre-exemples dans son analyse. Ainsi, nous rangeons les trois modalités de g_6 selon les deux ordres suivants :

Sc1 **: tolérance aux contre-exemples :** `conc` $\succ$ `lin` $\succ$ `conv` ;

Sc2 **: non tolérance aux contre-exemples :** `conv` $\succ$ `lin` $\succ$ `conc`.

On suppose de plus que l'on préfère une mesure dont la valeur est croissante avec n (l'ordre sur g_7 est alors `stat` $\succ$ `desc`), et qu'une lisibilité aisée de la mesure est souhaitable

	$g_1(\mu)$		$g_2(\mu)$		$g_3(\mu)$		$g_4(\mu)$		$g_5(\mu)$		$g_6(\mu)$			$g_7(\mu)$		$g_8(\mu)$		$g_9(\mu)$		
	sym	ori	non − dec_{p_b}	dec_{p_b}	max − var	max − fixe	indep − var	indep − fixe	indet − var	indet − fixe	conv	lin	conc	desc	stat	non − pc	pc	a	b	c
Sup	1	0	1	0	1	0	1	0	1	0	0	1	0	1	0	0	1	0	0	1
Conf	0	1	1	0	0	1	1	0	0	1	0	1	0	1	0	0	1	0	0	1
R	1	0	0	1	1	0	0	1	1	0	0	1	0	1	0	0	1	0	1	0
ConfCen	0	1	0	1	1	0	0	1	1	0	0	1	0	1	0	0	1	0	0	1
PS	1	0	0	1	1	0	0	1	1	0	0	1	0	1	0	0	1	0	1	0
Loe	0	1	0	1	0	1	0	1	1	0	0	1	0	1	0	0	1	0	1	0
Zhang	0	1	0	1	0	1	0	1	1	0	0	0	1	1	0	1	0	1	0	0
- IndImp	0	1	0	1	1	0	0	1	1	0	0	1	0	1	0	0	1	1	0	0
Lift	1	0	0	1	1	0	0	1	1	0	0	1	0	1	0	0	1	0	0	1
MoCo	0	1	1	0	1	0	1	0	0	1	0	1	0	1	0	0	1	0	1	0
Seb	0	1	1	0	0	1	1	0	0	1	1	0	0	1	0	0	1	0	1	0
FB	0	1	0	1	0	1	0	1	1	0	1	0	0	1	0	0	1	0	0	1
Conv	0	1	0	1	0	1	0	1	1	0	1	0	0	1	0	0	1	0	1	0
Tec	0	1	1	0	0	1	1	0	0	1	0	0	1	1	0	0	1	0	1	0
IQC	1	0	0	1	1	0	0	1	1	0	0	1	0	1	0	0	1	1	0	0
GI	1	0	0	1	1	0	0	1	1	0	0	0	1	1	0	0	1	1	0	0
IntImp	0	1	0	1	0	1	0	1	1	0	0	0	1	0	1	0	1	1	0	0
TIIE	0	1	0	1	0	1	0	1	0	1	0	0	1	0	1	1	0	1	0	0
IPD	0	1	0	1	1	0	0	1	1	0	0	0	1	0	1	0	1	1	0	0
Lap	0	1	1	0	1	0	1	0	1	0	0	1	0	1	0	0	1	1	0	0

Tab. III.27 – Matrice de synthèse des propriétés, sous forme normale disjonctive

(l'ordre sur g_9 est ainsi c $\succ$ b $\succ$ a).

La table III.28 donne les rangements totaux obtenus avec PROMETHEE.

Rang	1	2	3	4	5	6	7
Sc1 :	TIIE	IntImp	IPD	Loe	FB	Tec	Conf, ConfCen
Sc2 :	FB	Conv	Loe	Seb	ConfCen	Conf	TIIE
Rang	8	9	10	11	12	13	14
Sc1 :		Conv	Zhang	Lift	Seb	-IndImp	GI
Sc2 :	IntImp	Lift	IPD	-IndImp	Tec	MoCo	PS, r
Rang	15	16	17	18	19	20	
Sc1 :	PS, r		MoCo,	IQC	Sup	Lap	
Sc2 :		Zhang	IQC	Sup	GI	Lap	

TAB. III.28 – Rangements totaux associés aux scenarii Sc1 et Sc2

	Weight	Interval		% Weight	% Interval	
		Min	Max		Min	Max
g1	1.0000	0.0000	Infinity	11.11%	0.00%	100.00%
g2	1.0000	0.9500	Infinity	11.11%	10.61%	100.00%
g3	1.0000	0.7000	1.1000	11.11%	8.05%	12.09%
g4	1.0000	0.9500	Infinity	11.11%	10.61%	100.00%
g5	1.0000	1.0000	1.0500	11.11%	11.11%	11.60%
g6	1.0000	0.8824	1.0323	11.11%	9.93%	11.43%
g7	1.0000	0.9000	Infinity	11.11%	10.11%	100.00%
g8	1.0000	0.2000	1.0000	11.11%	2.44%	11.11%
g9	1.0000	0.9167	1.1333	11.11%	10.28%	12.41%

FIG. III.14 – Stabilité des poids pour Sc1

Les rangements obtenus sont bien dépendants du choix de l'utilisateur sur g_6. Dans Sc1, les trois mesures en tête de classement sont des mesures concaves. Elles sont également globalement bien évaluées par ailleurs, ce résultat n'est donc pas surprenant. Pour autant, Sc2 leur est défavorable, par la seule inversion de préférences sur g_6 : elles ne sont plus en tête du classement.

Deux autres mesures réaliseraient un bon compromis, car bien classées pour chacun des scenarii : FB et Loe.

PROMETHEE nous permet également de procéder à une analyse de stabilité. Cette analyse rend compte de l'écart qu'il est nécessaire d'apporter pour chaque poids afin de modifier les mesures en haut du classement (voire figures III.14 et III.15), dans notre cas, les cinq premières mesures. Dans le premier scénario, le jeu de poids sont relativement stables, et de légères fluctuations n'ont pas d'impact sur les cinq premières mesures proposées par la méthode. Le jeu de poids du second scénario semble toutefois plus sensible.

	Weight	Interval		% Weight	% Interval	
		Min	Max		Min	Max
g1	1.0000	0.0000	Infinity	11.11%	0.00%	100.00%
g2	1.0000	1.0000	1.1000	11.11%	11.11%	12.09%
g3	1.0000	0.9000	1.0000	11.11%	10.11%	11.11%
g4	1.0000	1.0000	1.1000	11.11%	11.11%	12.09%
g5	1.0000	0.9000	1.0000	11.11%	10.11%	11.11%
g6	1.0000	0.8571	1.4286	11.11%	9.68%	15.15%
g7	1.0000	0.0000	1.2000	11.11%	0.00%	13.04%
g8	1.0000	0.8000	Infinity	11.11%	9.09%	100.00%
g9	1.0000	0.8519	1.1667	11.11%	9.62%	12.73%

FIG. III.15 – Stabilité des poids pour SC2

Enfin, on peut visualiser l'axe de décision dans le plan GAIA. Cet axe, qui correspond à la projection du vecteur de décision est long dans SC1 (voire figure III.16). Des variations de préférences plus modérées auraient un faible impact sur les mesures situées dans la direction de cet axe. En revanche, pour SC2 (voire figure III.16), l'axe de décision est court. Ceci nous indique à nouveau que le rangement obtenu est plus sensible aux paramètres choisis.

Une mise en œuvre de ces résultats, dans un contexte applicatif industriel, est présenté dans [Plasse et al., 2006].

3 Agrégation de relations valuées

L'application d'une méthode d'AMD telle que celle que nous venons de présenter nous permet de sélectionner un sous-ensemble de mesures selon différents critères. Nous avons donc cherché à prendre en compte l'importance relative qu'apportent différentes mesures sur les règles [Barthélemy et al., 2006].

Pour ce faire, nous nous sommes intéressés à l'écart d'évaluation qui peut exister entre deux règles, pour une mesure donnée. Cet écart sera mesuré à l'aide d'une relation valuée de préférence R_j, pour chaque mesure μ_j, $j = 1, \ldots, m$.

Afin de procéder à la synthèse de ces différents avis, nous avons vu qu'il était nécessaire de nous munir d'une part d'une relation de préférence sur les couples de règles et d'autre part d'un opérateur d'agrégation défini sur ces relations.

3.1 Passage d'une mesure à une relation de préférence

Nous avons modélisé la relation de préférence entre deux règles au moyen de deux fonctions classiques de préférence.

La première fonction est linéaire par morceaux (voire figure III.18), et permet d'exprimer

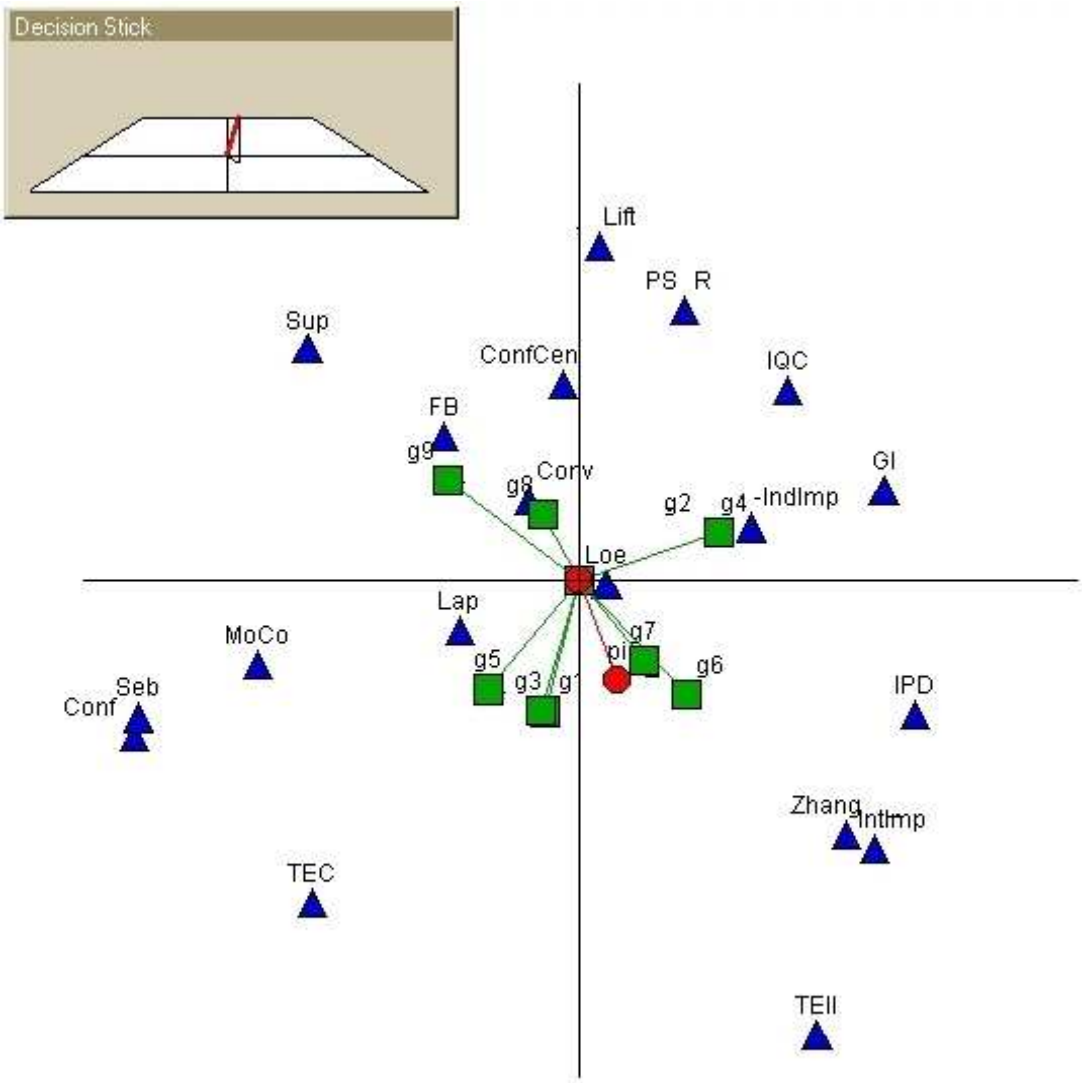


FIG. III.16 – Plan GAIA associé à SC1

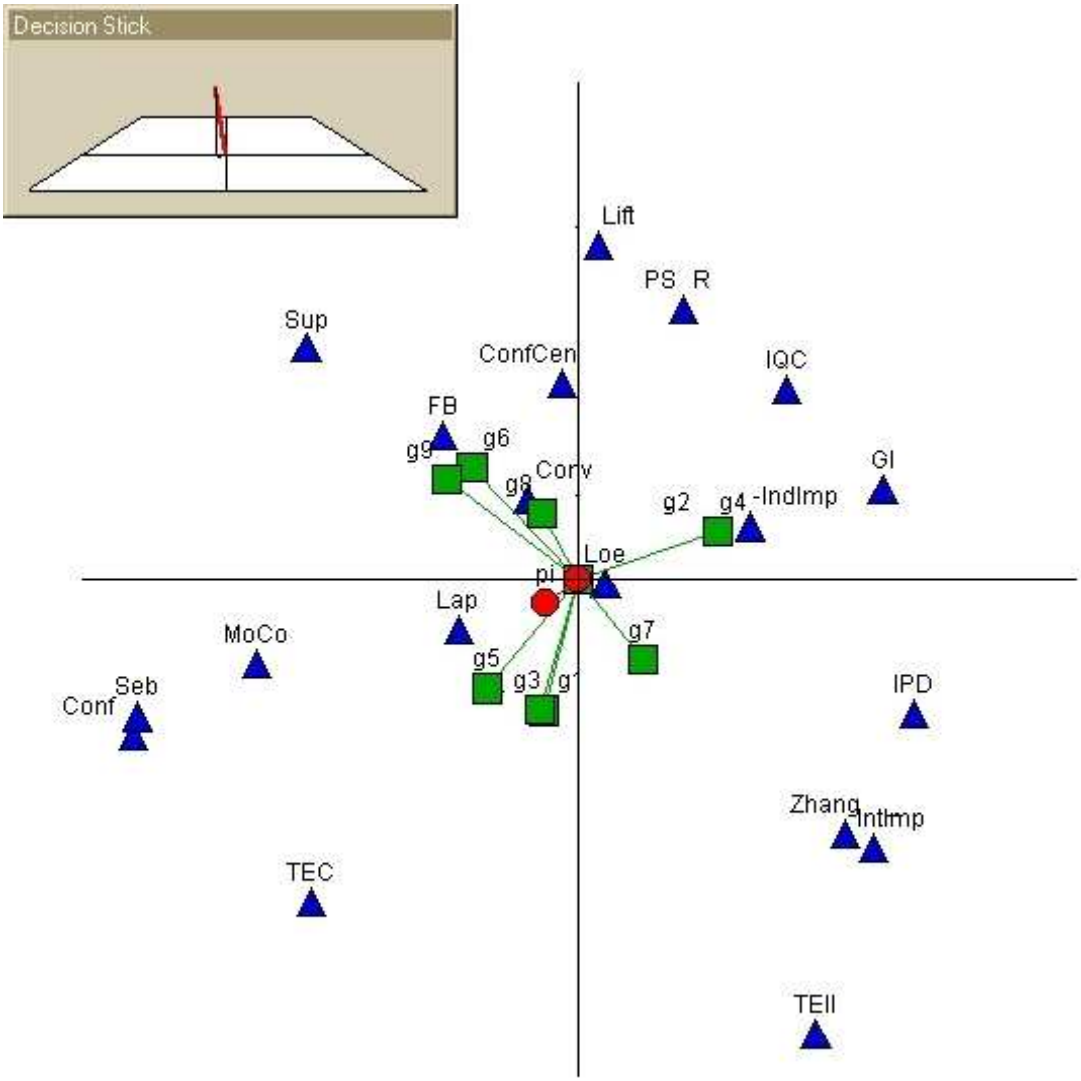


FIG. III.17 – Plan GAIA associé à SC2

que tout écart constant entre deux évaluations résulte en un écart de préférence proportionnel, ceci jusqu'à un certain seuil :

$$R_j(r_i, r_k) = \begin{cases} 0 & \text{si } \mu_j(r_i) - \mu_j(r_k) < 0 \\ \frac{\mu_j(r_i) - \mu_j(r_k)}{p_j} & \text{si } 0 \leq \mu_j(r_i) - \mu_j(r_k) \leq p_j \\ 1 & \text{si } \mu_j(r_i) - \mu_j(r_k) > p_j \end{cases}$$

où p_j correspond au seuil de préférence à partir duquel on considère que l'intensité de préférence est "très forte" (égale à 1). Cependant, bien qu'elles présentent un certain intérêt de par la simplicité de leur expression et leurs propriétés naturelles, notamment le respect des propriétés d'irréflexivité, de min-transitivité et de min-asymétrie, nous n'avons pas retenu les représentations de préférences individuelles de ce type dans notre étude. Leur linéarité ne permet pas de traduire les préférences qui sont exercées entre les règles de façon très pertinente. En effet, le passage de la non préférence ($R_j(r_i, r_k) = 0$) à la préférence faible ($R_j(r_i, r_k) = \frac{\mu_j(r_i) - \mu_j(r_k)}{p_j}$) et de la préférence faible à la préférence forte ($R_j(r_i, r_k) = 1$) s'effectue de manière "brutale" ce qui se traduit par l'existence de deux points de ruptures d'abscisses $\mu_j(r_i) - \mu_j(r_k) = 0$ et p_j.

Ainsi, nous nous sommes intéressés à une relation de préférence moins abrupte. Nous avons retenu celle proposée par [Brans et al., 1984], classiquement utilisée en modélisation de préférences, lorsque l'on souhaite justement adoucir les phases de transitions entre les différentes situations d'indifférence, de préférence faible et de préférence forte.

Cette relation de préférence (voire figure III.19), appelée indice gaussien, permet de traduire une préférence non linéaire d'une règle règle r_i sur une règle r_k. On la définit comme suit :

$$R_j(r_i, r_k) = \begin{cases} 1 - exp\left(-\frac{(\mu_j(r_i) - \mu_j(r_k))^2}{2\sigma_j^2}\right) & \text{si } \mu_j(r_i) - \mu_j(r_k) > 0 \\ 0 & \text{sinon} \end{cases}$$

où σ_j est un paramètre qui représente à la fois la position du point d'inflexion de la fonction (la dérivée seconde de la fonction s'annulant pour $\mu_j(r_i) - \mu_j(r_k) = \sigma_j$) mais aussi une valeur de seuil entre les zones de préférence faible et de préférence forte. Nous discuterons plus loin de la valeur à affecter à ce paramètre.

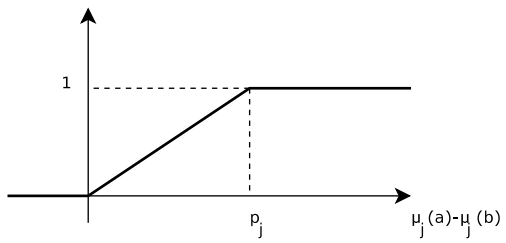


FIG. III.18 – Fonction de préférence linéaire

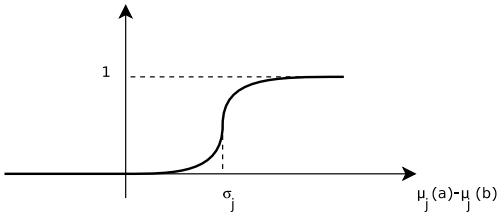


FIG. III.19 – fonction de préférence gaussienne

On montre dans [Legrain, 2004] que la relation de préférence issue de l'indice gaussien est min-transitive.

3.2 Choix d'opérateur d'agrégation

Nous avons présenté en section 1.2.4 différents opérateurs d'agrégation classiques. Parmi ces opérateurs, nous avons examiné ceux étant susceptibles de répondre à un certain nombre d'exigences.

Les moyennes d'ordre α, avec $\alpha < 0$ et la moyenne géométrique seront soumises à des effets de veto dès lors que deux règles distinctes seront évaluées également par l'une des mesures. En effet, la présence d'une valeur nulle associée à une différence d'évaluation nulle entre deux règles "écrasera" numériquement les autres différences d'évaluations.

On peut donc penser qu'il soit plus juste de s'orienter vers des moyennes d'ordre α, avec $\alpha > 0$. Toutefois, celles-ci présentent l'inconvénient, partagé avec les opérateurs précédents, de ne pas conserver la min-transitivité (comme résumé dans le tableau III.1).

D'autres opérateurs plus complexes auraient pu être envisagés. Parmi les opérateurs classiques, on aurait pu par exemple s'orienter vers les intégrales floues telles celle de Choquet [Choquet, 1953], ou celle de Sugeno [Sugeno, 1974], mais celles-ci ont le désavantage d'être sensibles aux coalitions (phénomène que nous souhaitons éviter). Or, ces situations sont loin d'être exclues, ne serait-ce que parce que certaines mesures vont ordonner les règles de manière très semblable, comme nous le verrons au chapitre IV. Quand aux opérateurs OWA [Yager, 1988] ou au minimum et maximum pondérés, ceux-ci n'opèrent pas sur les grandeurs qui nous intéressent. La première famille d'opérateurs travaille sur le rangs, et ne prend pas en compte la différence d'évaluation numérique. Les dernières ne sont pas compensatoires : elles favorisent l'une des préférences, ignorant les autres.

Nous allons illustrer les résultats obtenus par agrégation en appliquant notre procédure au cas d'école, que nous avons introduit en section 2, et qui contient un faible nombre de règles.

L'évaluation des règles par les mesures se fait au moyen du logiciel HERBS (voire chapitre IV). Puis, les valeurs numériques issues des évaluations sont agrégées au moyen d'un programme python.

3.3 Etude de cas

Nous disposons de la base de données transactionnelle décrite dans le tableau I.1. Nous avons appliqué l'algorithme APRIORI en fixant un seuil de support à 20%, un seuil de confiance de 60% et un nombre d'attributs compris entre 2 et 5 par itemset généré. Nous avons ainsi obtenons les 21 règles listées dans le tableau I.3.

Le tableau II.5 liste la valeur de chaque mesure de qualité pour chacune des règles. Les préordres générés par les mesures de qualité sont variés en nature (taille et nombre de groupes de règles), et en contenu (variabilité des ordonnancements des règles). Ils sont listés dans le tableau II.6. On rappelle que trois règles y sont mises en valeur :

r_{18} est une règle qui est globalement bien évaluée par toutes les mesures. Il est raisonnable de globalement considérer qu'il s'agit d'une "bonne" règle.

r_{19} à l'inverse est constamment en fin de préordre. De manière similaire, il est logique de la considérer globalement comme mauvaise.

r_2 par contre n'est ni toujours bonne, ni toujours mauvaise, et illustre notre problématique. Selon certaines mesures comme Lift, GI ou encore R, elle est très bonne, mais selon d'autres mesures comme Conv ou MoCo elle est très mauvaise !

A partir des différences d'évaluations, on peut calculer les valeurs des préférences individuelles pour chaque paire de règles. Pour cela nous rappelons la fonction que nous utilisons :

$$R_j(r_i, r_k) = \begin{cases} 1 - exp\left(-\frac{(\mu_j(r_i)-\mu_j(r_k))^2}{2\sigma_j^2}\right) & \text{si } \mu_j(r_i) - \mu_j(r_k) > 0 \\ 0 & \text{sinon} \end{cases}$$

3.3.1 Construction de relations de préférences floues

Le seuil σ_j est fixé de la façon suivante : pour chaque mesure μ_j et toutes règles r_i, r_k, $i, k \in [\![1, n]\!]$, nous calculons les $\frac{n(n-1)}{2}$ écarts de valeurs $|\mu_j(r_i) - \mu_j(r_k)|$ et les ordonnons par ordre croissant. On sélectionne ensuite la valeur de σ_j à une proportion donnée de ces écarts. Par exemple, un seuil σ_j fixé à 50% correspond alors à la médiane des différences. La table III.29 récapitule différentes valeurs de σ_j que nous allons prendre afin de calculer les intensités de préférence.

Nous obtenons ainsi 20 matrices donnant les degrés d'intensité de préférence entre toutes les paires de règles. Nous ne les présentons pas car elles n'offrent que peu d'intérêt. En revanche ces informations sont contenues dans le tableau III.31 qui les met en lien avec les relations agrégées.

Au moyen de la moyenne arithmétique, un jeu de poids équilibré, et des seuils σ_j pris pour un écart de 40%, l'opération d'agrégation permet d'obtenir les résultats présentés en tables III.30 et III.31. Le premier tableau correspond à la matrice des préférences globales valuées et le second met en relation les préférences individuelles obtenues sur chaque mesure et les résultats de l'agrégation. Le rayon du disque traduit la valeur de la préférence globale entre chaque couple de règles, un rayon de 1 étant attribué à une préférence globale égale à 1, et un rayon de $0,5$ à une préférence globale nulle. L'intérieur du cercle permet de représenter les valeurs des préférences individuelles pour chacune des 20 mesures, un dégradé de couleur allant du rouge au bleu permet de représenter les intensités de la plus faible à la plus forte.

Afin d'interpréter les résultats obtenus en table III.31, on procède de la façon suivante : la ligne qui correspond à une règle permet de visualiser l'intensité de préférence que cette règle exerce sur les autres. En prenant l'exemple de r_{18} qui était souvent bien classé par chacune des mesures (les subdivisions à l'intérieur des cercles sont majoritairement de couleur bleue), on remarque, de par les rayons des cercles qui sont proches de 1 sauf pour ceux qui la comparent à r_1 et r_{17}, que la préférence globale de cette règle sur les autres est souvent élevée, ce qui nous permet de dire qu'elle est une "bonne" règle. Les résultats face à r_1 et r_{17} ne sont pas surprenant car ces deux règles sont elles aussi souvent bien classées par les mesures.

	40%	46%	53%	60%	66%	écart maximal
-IndImp	0.51	0.61	0.73	0.73	0.8	1.95
ConfCen	0.15	0.2	0.2	0.25	0.28	0.63
Conf	0.08	0.13	0.13	0.15	0.2	0.4
Conv	0.6	0.8	1.0	1.4	2.2	$+\infty$
r	0.1	0.12	0.13	0.13	0.16	0.38
FB	1.25	1.67	2.17	3.5	6.5	$+\infty$
GI	0.1	0.12	0.13	0.18	0.18	0.48
IntImp	0.18	0.22	0.26	0.27	0.28	0.66
IPD	0.26	0.28	0.33	0.43	0.47	0.9
IQC	0.23	0.36	0.38	0.4	0.43	1.14
Lap	0.05	0.09	0.09	0.1	0.1	0.23
Lift	0.25	0.31	0.39	0.5	0.5	1.67
Loe	6.0	6.67	8.0	9.33	10.5	28.0
PS	0.06	0.08	0.09	0.1	0.11	0.26
Seb	1.0	1.5	2.0	2.0	2.5	$+\infty$
Sup	0.0	0.0	0.1	0.1	0.1	0.1
MoCo	0.08	0.13	0.15	0.16	0.21	0.5
Tec	0.17	0.25	0.25	0.33	0.33	0.67
TIIE	0.24	0.46	0.51	0.63	0.7	0.94
Zhang	0.5	0.5	0.5	0.71	0.75	1.5

Tab. III.29 – Différentes valeurs de σ

En observant les colonnes, nous obtenons ce qu'on pourrait appeler le degré de non préférence de toutes les autres règles sur la règle considérée. Ainsi la colonne r_{19} possède plusieurs cercles aux rayons proches de 1, ce qui est un résultat attendu étant donné le mauvais classement qu'elle obtient sur les différentes mesures, mais aussi des cercles à petit rayon traduisant une "non-préférence" globale proche de 0. Ce résultat est quelque peu inattendu.

Enfin, la règle r_2 qui se "baladait" dans les différents préordres induits par les mesures révèle des intensités de préférence variables selon les règles auxquelles elle est comparée, ce qui correspond au résultat attendu.

3.3.2 Retour à une relation de préférence nette

Afin de mieux visualiser les relations de dominance entre règles, on peut souhaiter revenir à une relation de préférence nette. Classiquement, on revient à une relation nette à partir de la relation floue en fixant un seuil minimal au-delà duquel une règle sera préférée à une autre.

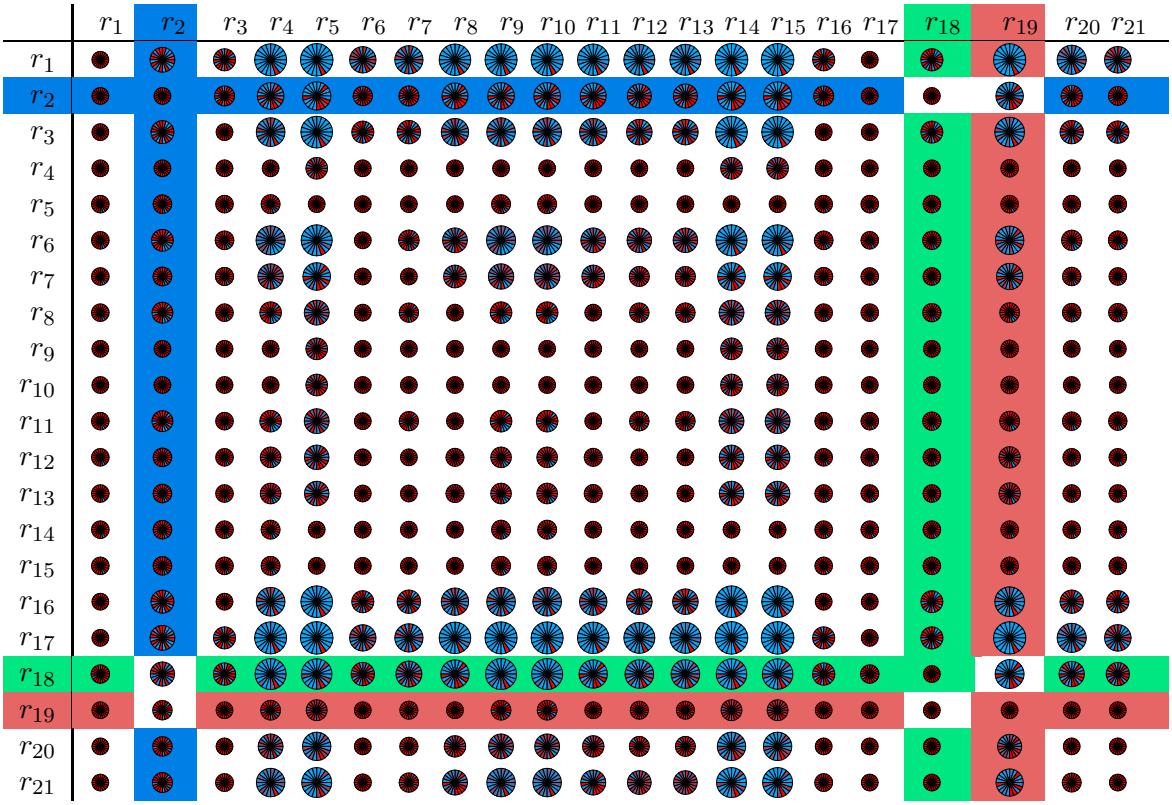
Définition 26 (coupe λ) *Pour tout $\lambda \in [0,1]$, la relation R_λ sur $\mathcal{R}^2$, appelée coupe λ de R, est définie comme l'ensemble des paires de $\mathcal{R} \times \mathcal{R}$ dont les valeurs ne sont pas plus mauvaises que λ :*

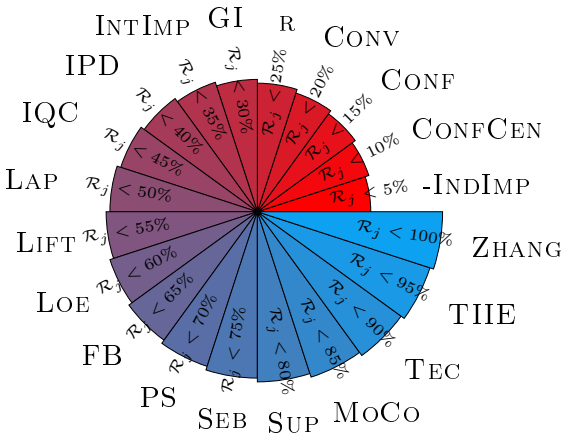
$$R_\lambda = \{(r_i, r_k) \in \mathcal{R} \times \mathcal{R} | R(r_i, r_k) \geq \lambda\}$$

	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}	r_{11}	r_{12}	r_{13}	r_{14}	r_{15}	r_{16}	r_{17}	r_{18}	r_{19}	r_{20}	r_{21}
r_1	0	0.28	0.32	0.94	1	0.39	0.54	0.94	0.94	0.94	0.94	0.83	0.83	1	1	0.32	0	0.04	0.98	0.65	0.39
r_2	0	0	0	0.7	0.99	0.02	0.09	0.7	0.7	0.7	0.7	0.43	0.43	0.99	0.99	0	0	0	0.86	0.18	0.02
r_3	0	0	0	0.68	0.99	0.01	0.07	0.68	0.68	0.68	0.68	0.39	0.39	0.99	0.99	0	0	0	0.85	0.16	0.01
r_4	0	0	0	0	0.63	0	0	0	0	0	0	0	0	0.63	0.63	0	0	0	0.09	0	0
r_5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r_6	0	0	0	0.61	0.98	0	0.03	0.61	0.61	0.61	0.61	0.32	0.32	0.98	0.98	0	0	0	0.81	0.1	0
r_7	0	0	0	0.47	0.96	0	0	0.47	0.47	0.47	0.47	0.17	0.17	0.96	0.96	0	0	0	0.7	0.02	0
r_8	0	0	0	0	0.63	0	0	0	0	0	0	0	0	0.63	0.63	0	0	0	0.09	0	0
r_9	0	0	0	0	0.63	0	0	0	0	0	0	0	0	0.63	0.63	0	0	0	0.09	0	0
r_{10}	0	0	0	0	0.63	0	0	0	0	0	0	0	0	0.63	0.63	0	0	0	0.09	0	0
r_{11}	0	0	0	0	0.63	0	0	0	0	0	0	0	0	0.63	0.63	0	0	0	0.09	0	0
r_{12}	0	0	0	0.12	0.84	0	0	0.12	0.12	0.12	0.12	0	0	0.84	0.84	0	0	0	0.36	0	0
r_{13}	0	0	0	0.12	0.84	0	0	0.12	0.12	0.12	0.12	0	0	0.84	0.84	0	0	0	0.36	0	0
r_{14}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r_{15}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r_{16}	0	0	0	0.68	0.99	0.01	0.07	0.68	0.68	0.68	0.68	0.39	0.39	0.99	0.99	0	0	0	0.85	0.16	0.01
r_{17}	0	0.28	0.32	0.94	1	0.39	0.54	0.94	0.94	0.94	0.94	0.83	0.83	1	1	0.32	0	0.04	0.98	0.65	0.39
r_{18}	0	0.13	0.16	0.89	1	0.22	0.37	0.89	0.89	0.89	0.89	0.71	0.71	1	1	0.16	0	0	0.96	0.49	0.22
r_{19}	0	0	0	0	0.38	0	0	0	0	0	0	0	0	0.38	0.38	0	0	0	0	0	0
r_{20}	0	0	0	0.35	0.93	0	0	0.35	0.35	0.35	0.35	0.08	0.08	0.93	0.93	0	0	0	0.6	0	0
r_{21}	0	0	0	0.61	0.98	0	0.03	0.61	0.61	0.61	0.61	0.32	0.32	0.98	0.98	0	0	0	0.81	0.1	0

TAB. III.30 – Agrégation des préférences (exemple jouet, σ pris à 40%, poids équilibrés)

Ainsi, en adoptant un seuil de coupe λ approprié nous allons obtenir un graphe représentant uniquement les relations permettant de dire qu'une règle est "globalement meilleure" qu'une autre.





Tab. III.31 – Agrégation de préférences

La figure III.20 représente le graphe complet de la relation globale ; toutes les règles sont liées les unes aux autres. Ce graphe n'a pas d'intérêt réel, étant donné que nous nous intéressons aux règles ayant une intensité de préférence globale "remarquable". Il permet cependant d'illustrer la complexité de la relation sous-jacente. C'est pourquoi nous allons procéder à des coupes des relations valuées par la suite.

Afin de lire ces figures, il convient d'identifier les arêtes entre les règles, et non leurs positions relatives. Lorsqu'une règle r_i est préférée à une règle r_k, une arête les relie, et la règle r_i est placée à gauche. S'il n'existe pas de lien reliant deux règles, leurs positions relatives dans le graphe n'est due qu'à des besoins de représentations. Les graphes ont été générés au moyen de l'outil DOT, de la suite GRAPHVIZ[4].

Les graphes III.21 à III.25 ont tous été obtenus à l'aide d'un seuil de coupe de 0.4. On observe qu'en augmentant la valeur de σ_j, le nombre d'arêtes diminue, ce qui traduit des préférences globales plus faibles qu'auparavant. Ceci s'explique par le fait que σ_j représente un indicateur d'étendue. Plus il est proche de 0, plus la relation valuée se comporte comme une relation nette. Pour une valeur de σ_j importante, pris à un écart de 66%, les relations forment un graphe biparti : seul un écart de valeur important sur une mesure sera considéré comme relevant d'une préférence individuelle entre deux règles.

Si l'on observe les résultats obtenus en faisant maintenant varier λ (figures III.26 à III.30), avec des seuils σ_j à nouveau pris à 40%, nous observons que plus λ est grand, plus le nombre d'arêtes est faible : le seuil minimal de préférence est plus fort. Avec $\lambda = 0.1$, le graphe reste peu lisible, mais nous pouvons remarquer un résultat surprenant : r_2 est bien classée. Ceci s'explique simplement par le fait que le seuil de coupure est si bas, que presque n'importe quelle règle est préférée à toutes les autres. Dans cette situation, DOT est incapable de placer correctement les nœuds du graphe, et adopte des choix de placement "ad hoc", en l'absence de choix utilisateur. Les nœuds sont placés dans leur ordre de description, ce qui explique que r_2 se retrouve ainsi incorrectement positionnée. Pour un seuil λ de 0.4, on obtient un graphe ne contenant plus de situations conflictuelles.

Ces situations conflictuelles que l'on observe ici étaient envisagées, puisque nous avons vu que l'ensemble des propriétés dont nous souhaitions disposer ne peuvent être retenues simultanément. De plus, la petite taille du jeu de données utilisé favorise l'apparition de situations de veto ou d'additivité.

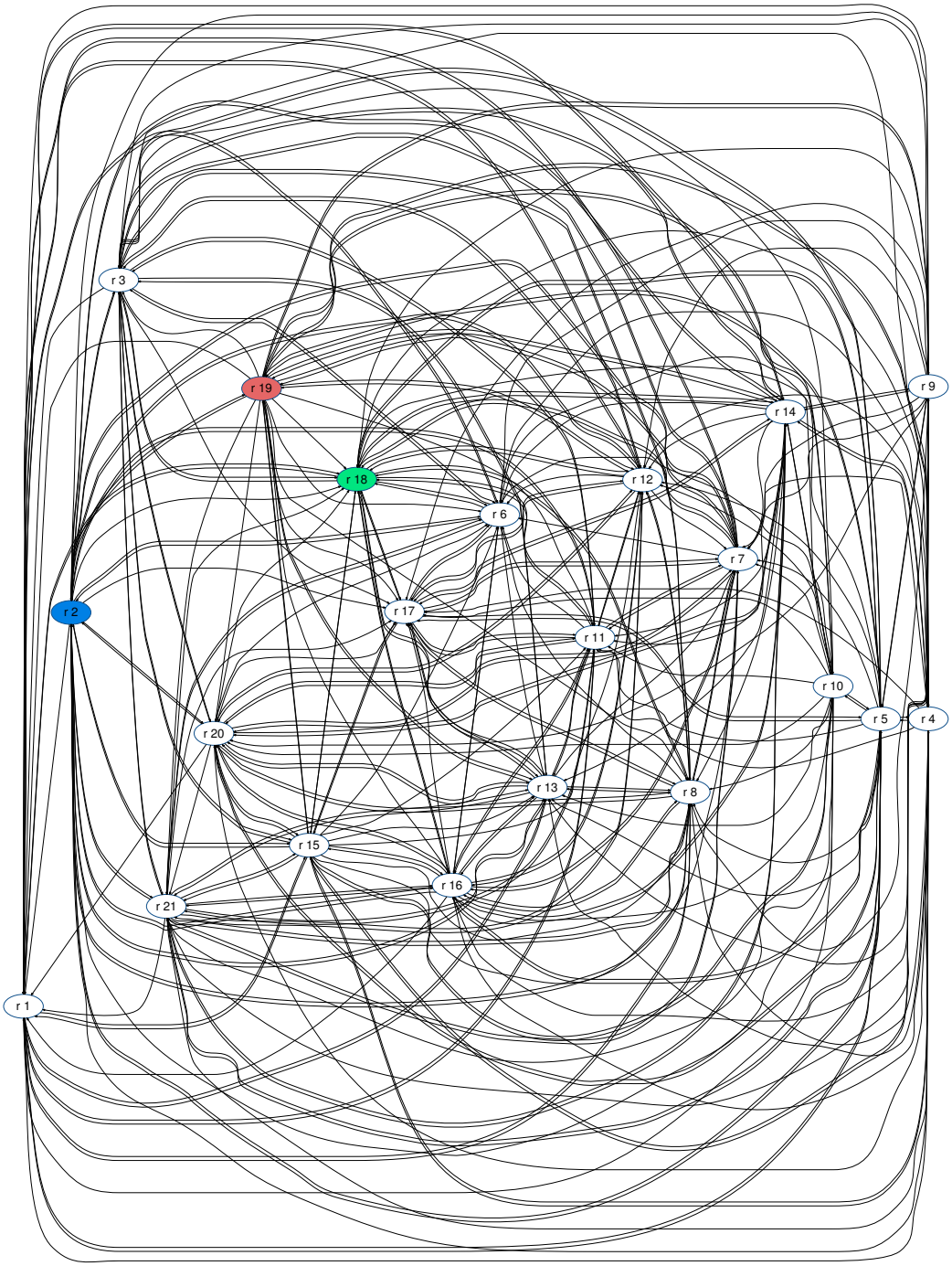


FIG. III.20 – Graphe de préférences pour un seuil de coupure nul

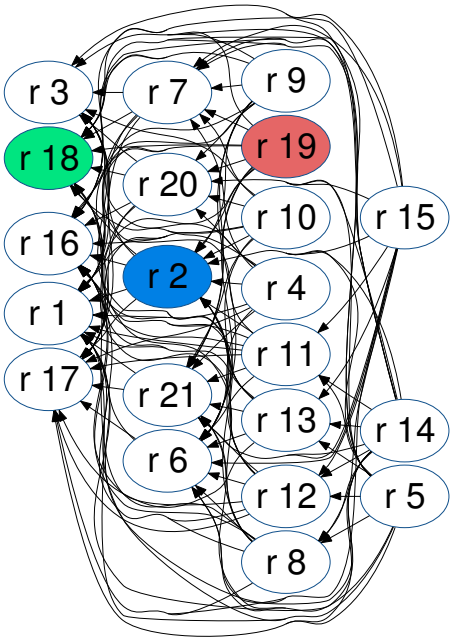


FIG. III.21 – Graphe de préférences : σ_j pris à un écart de 40%

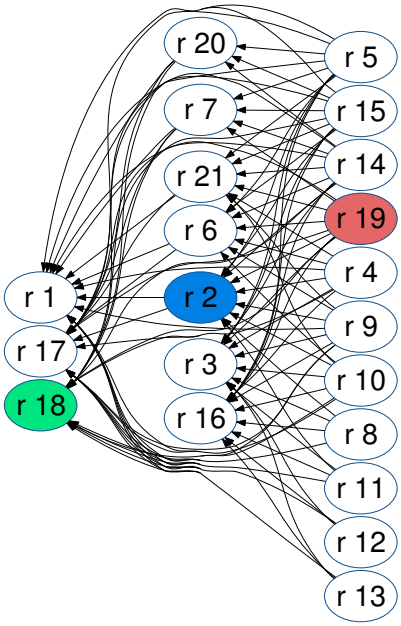


FIG. III.22 – Graphe de préférences : σ_j pris à un écart de 46%

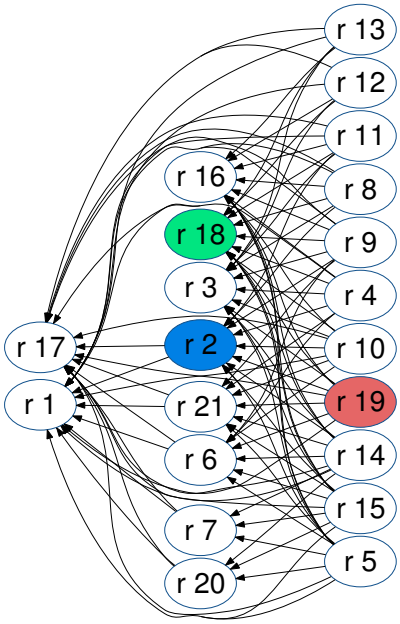


FIG. III.23 – Graphe de préférences : σ_j pris à un écart de 53%

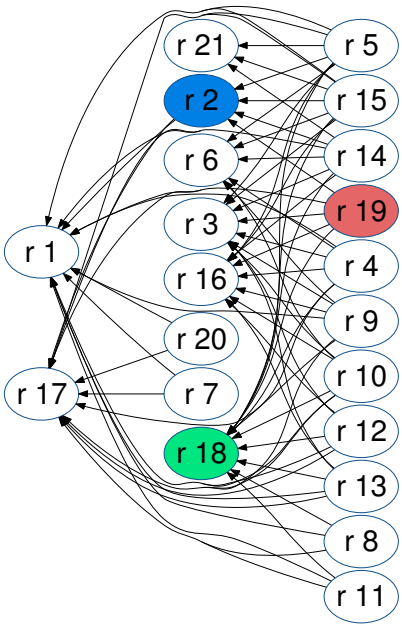


FIG. III.24 – Graphe de préférences : σ_j pris à un écart de 60%

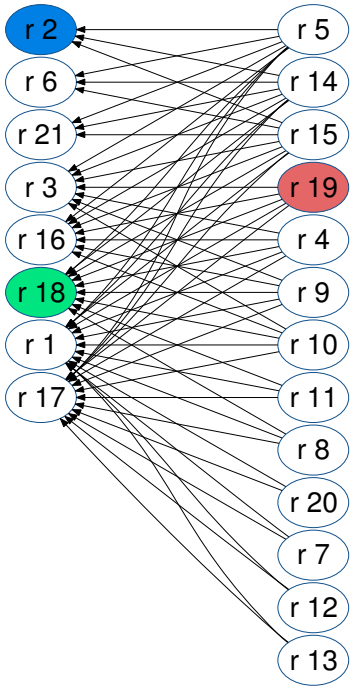


FIG. III.25 – Graphe de préférences : σ_j pris à un écart de 66%

Afin de tenter de minimiser d'éventuels effets de coalition, nous pouvons observer le résultat produit par l'agrégation en utilisant un jeu de poids non homogène. Nous allons construire deux jeux de poids, à partir des résultats obtenus en section 2. Reprenant les rangements des mesures obtenus via la méthode PROMETHEE pour les scénario SC1 et SC2, nous assignons à chaque mesure, et après normalisation, un poids inverse à son rang. Ainsi, les premières mesures du classement auront une contribution plus forte que les dernières. Les deux jeux de poids associés à chaque scénario sont listés en table III.32.

Une alternative aurait été de ne retenir qu'un sous-groupe de mesures. Dans sa dernière version, PROMETHEE permet en effet de ne sélectionner, parmi les alternatives possibles, qu'un sous-ensemble d'entre elles, ce choix étant composé –par exemple– de quelques "bonnes" alternatives, quelques "moyennes", et quelques "mauvaises".

Les résultats obtenus sont plus intéressants que les précédents. Dans le cas SC1 (figures III.35 à III.38), la règle r_2 est située au même niveau que r_{18}, même en faisant varier σ_j, avec $\lambda = 0.9$. Pour SC2 (figures III.35 à III.38), ils paraissent plus concluants car respectant mieux l'ordre sur nos trois règles de référence. En prenant des seuils σ_j à 5% et des seuils λ relativement élevés (0.8 et 0.9) nous obtenons un graphe qui classe les règles en respectant les préordres sur chaque mesure.

[4] http ://www.graphviz.org/

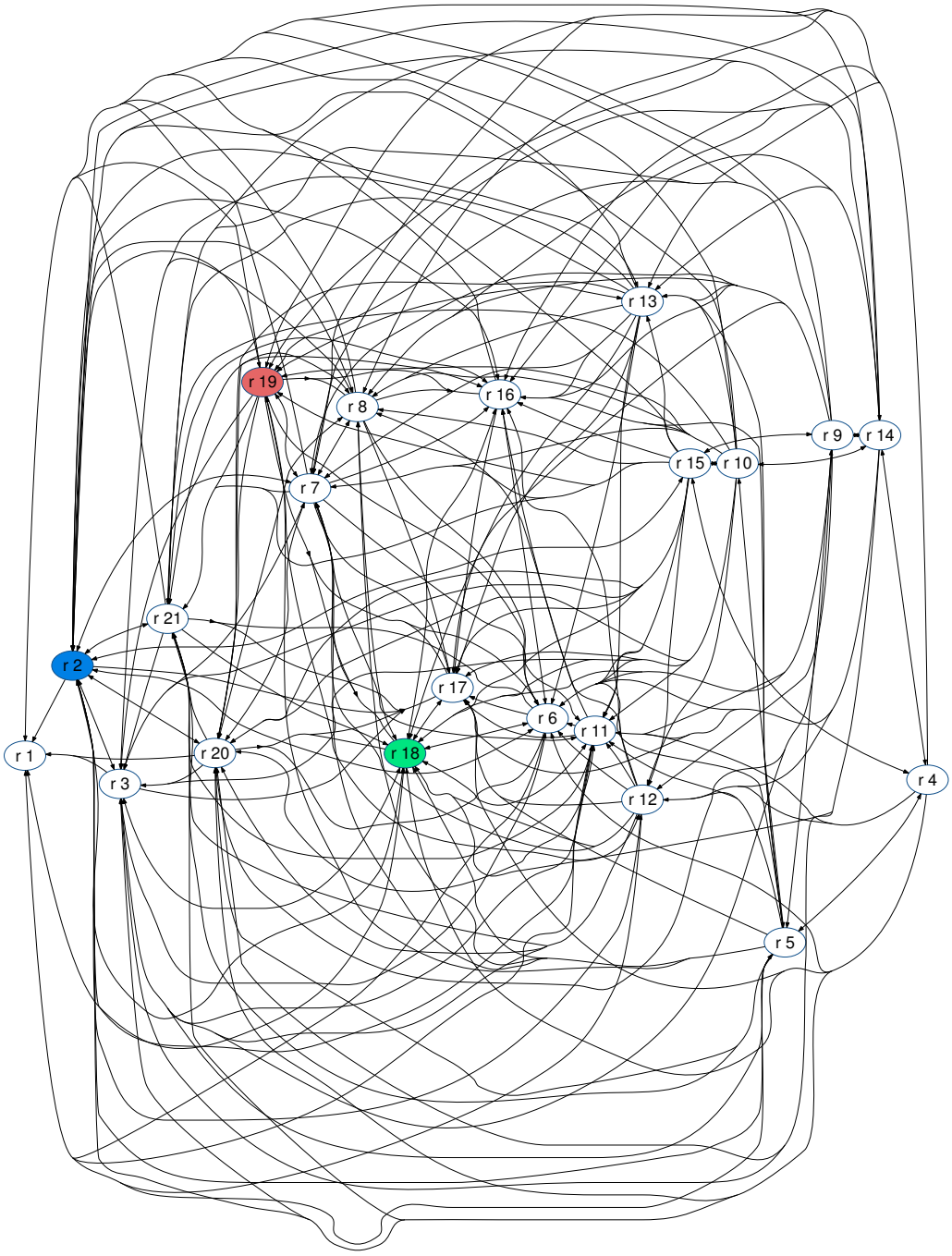


FIG. III.26 – Graphe de préférence pour $\lambda = 0.1$

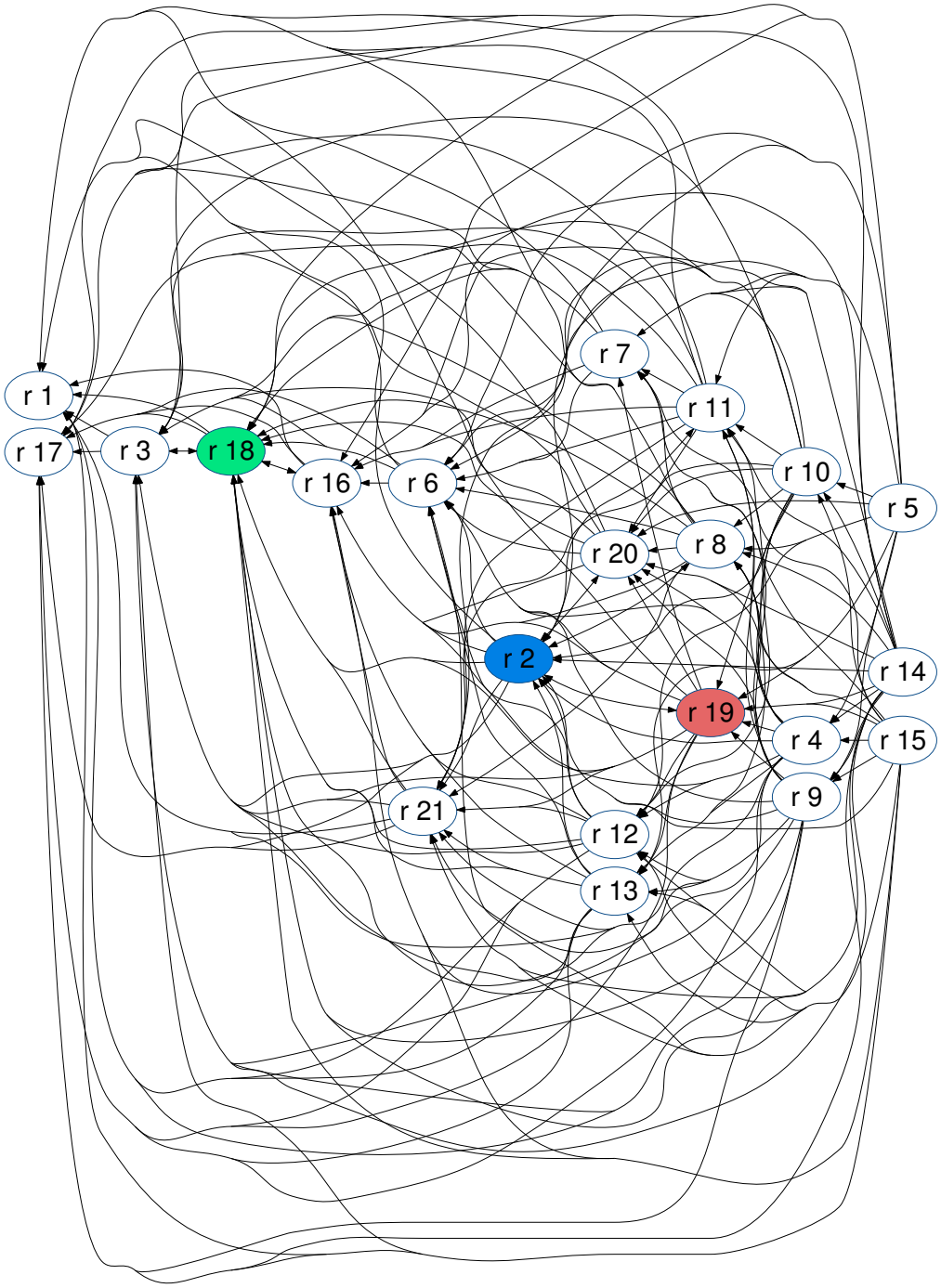


FIG. III.27 – Graphe de préférence pour $\lambda = 0.2$

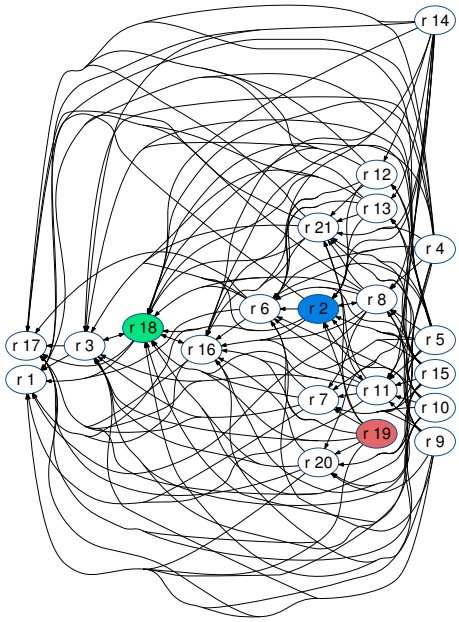


FIG. III.28 – Graphe de préférence pour $\lambda = 0.3$

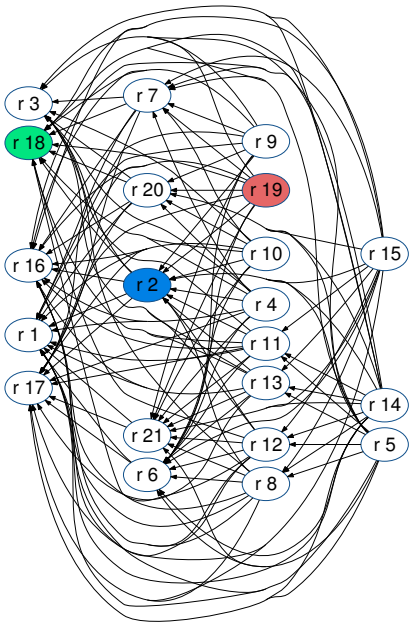


FIG. III.29 – Graphe de préférence pour $\lambda = 0.4$

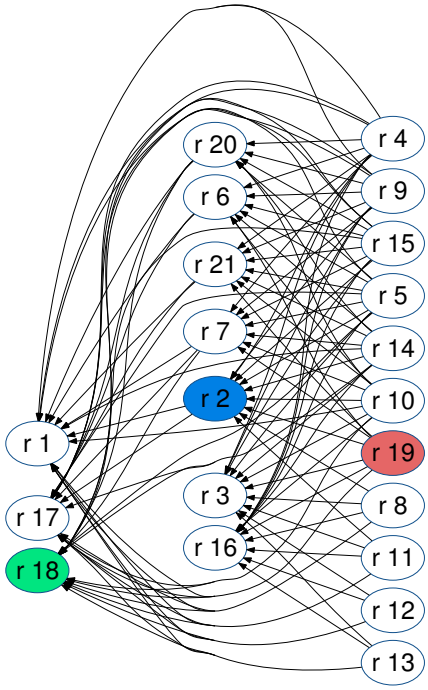


FIG. III.30 – Graphe de préférence pour $\lambda = 0.5$

mesure	poids associé à Sc1	poids associé à Sc2
-IndImp	0.0395	0.0459
ConfCen	0.0553	0.0689
Conf	0.0230	0.0197
Conv	0.0461	0.1378
r	0.0198	0.0250
GI	0.0213	0.0162
IntImp	0.2766	0.0551
TIIE	0.1383	0.0344
IPD	0.0395	0.0459
IQC	0.0154	0.0184
Lap	0.0138	0.0145
Lift	0.0184	0.0250
Loe	0.0922	0.0918
FB	0.0691	0.2755
PS	0.0277	0.0276
Seb	0.0163	0.0212
Sup	0.0146	0.0138
MoCo	0.0173	0.0172
Tec	0.0251	0.0153
Zhang	0.0307	0.0306

Tab. III.32 – Jeu de poids utilisés dans les scénario Sc1 et Sc2

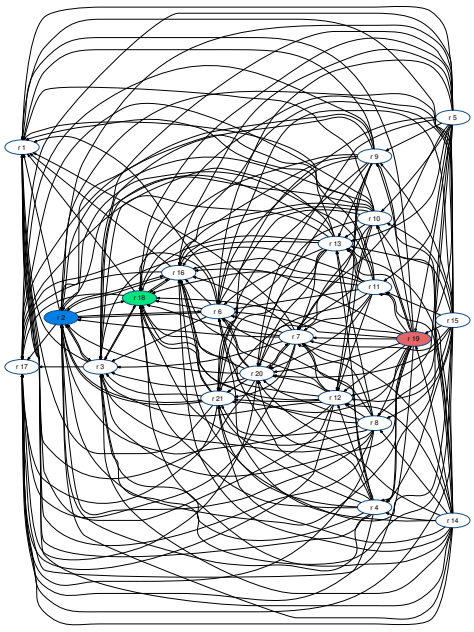


Fig. III.31 – Jeu de poids Sc1, σ_j pris à un écart de 5%

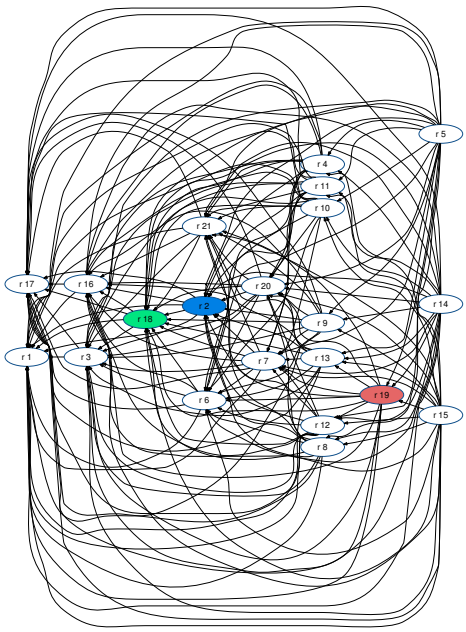


Fig. III.32 – Jeu de poids Sc1, σ_j pris à un écart de 25%

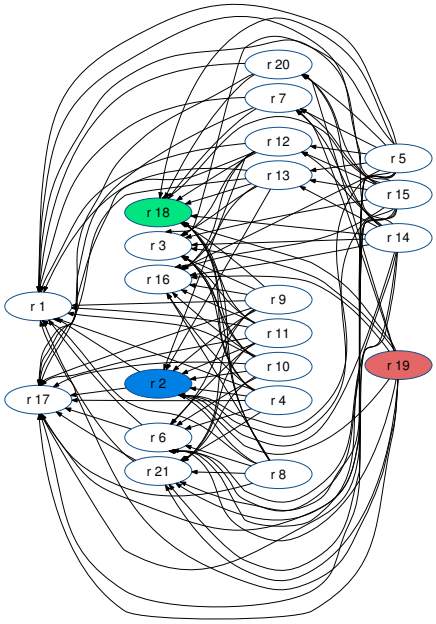


FIG. III.33 – Jeu de poids SC1, σ_j pris à un écart de 50%

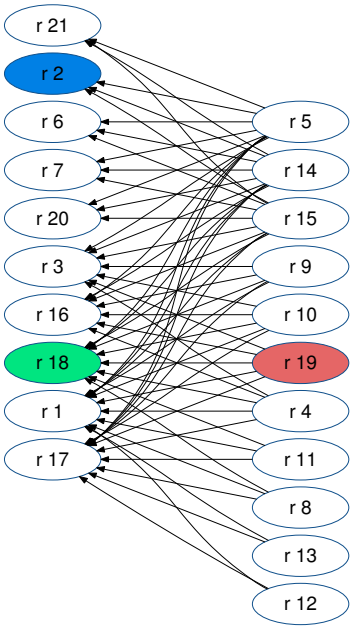


FIG. III.34 – Jeu de poids SC1, σ_j pris à un écart de 75%

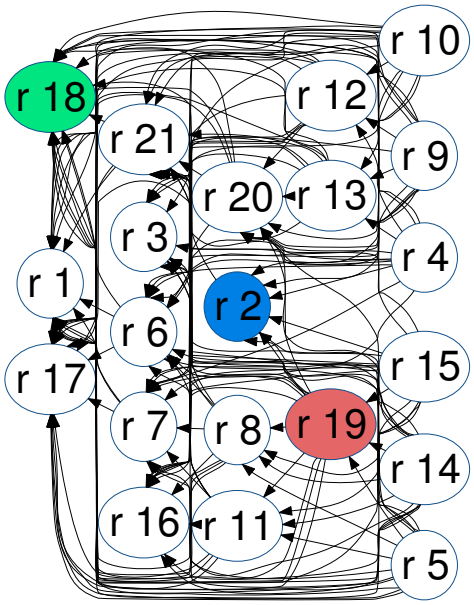


FIG. III.35 – Jeu de poids SC2, $\lambda = 0,9$

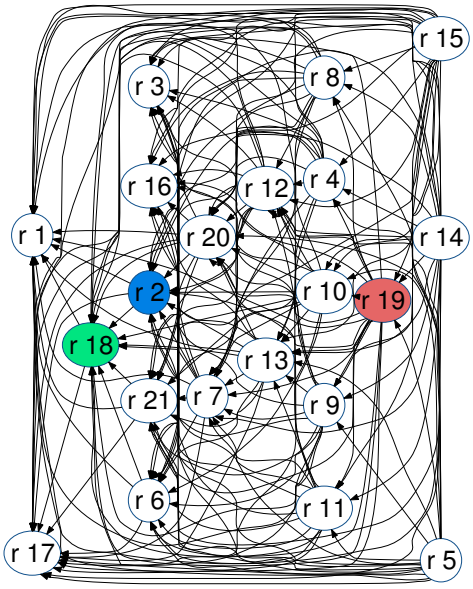


FIG. III.36 – Jeu de poids SC2, $\lambda = 0,8$

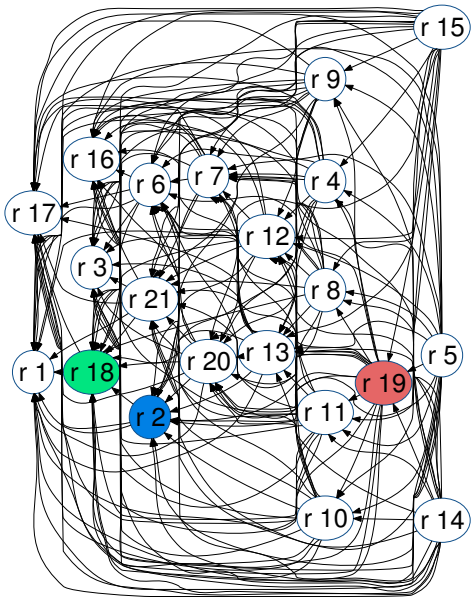


FIG. III.37 – Jeu de poids SC2, $\lambda = 0,7$

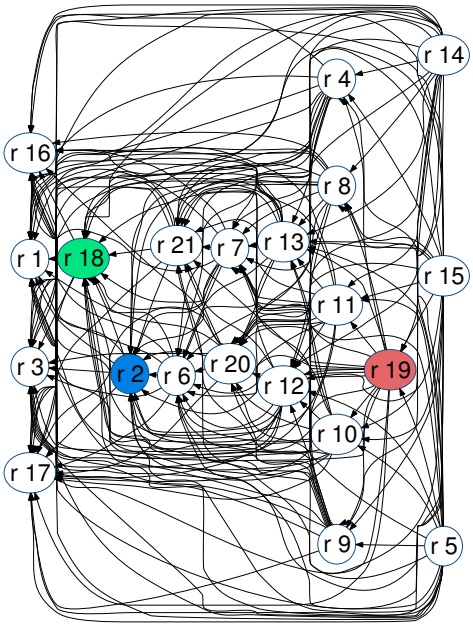


FIG. III.38 – Jeu de poids SC2, $\lambda = 0,6$

Des expériences sur de plus grandes bases de règles ont été effectuées, par exemple dans [Legrain, 2004]. Nous ne les présentons pas dans ce mémoire car la méthode de représentation visuelle automatique via DOT ne permet pas d'obtenir de résultat visuel satisfaisant, principalement à cause du trop grand nombre d'arêtes à placer. Une voie à explorer pour résoudre ce problème de représentation serait d'offrir la possibilité de ne focaliser que sur des sous-ensembles de règles, afin de procéder à une exploration locale de la base de règles, et ainsi de limiter la taille du graphe de dominance affiché. Cette approche nécessite des outils de visualisation et de navigation avancés, ainsi qu'une coopération efficace avec l'utilisateur. Ce type d'approche prometteuse a été développé en fouille interactive de donnée, par exemple pour l'exploration de bases de règles d'association [Lehn, 2000, Kuntz et al., 2004]. Il est également possible de s'orienter vers une recherche de compromis [Francisci et al., 2003].

Chapitre IV

Expérimentations

Afin de compléter nos études formelles, nous avons mené parallèlement des études "en situation", c'est-à-dire sur des données. De plus, bien que les propriétés que nous avons étudié couvrent en grande partie la plupart des propriétés proposées dans la littérature, nous ne pouvons prétendre à l'exhaustivité. Qui plus est, des spécificités dans des jeux de données pourraient engendrer des résultats imprévus en théorie. Par ailleurs, à notre connaissance, aucun outil n'intégrait lors des premières campagnes expérimentales que nous avons conduit un ensemble conséquent de mesures de qualité. Il était particulièrement marquant de voir que de nombreux logiciels ne proposaient, dans le cadre des règles d'association, que SUP, CONF et LIFT, ceci malgré le foisonnement des mesures de qualité.

Ces considérations nous ont mené à développer l'outil HERBS [Vaillant, 2002], dédié à la post-analyse. Les bases de cas et de règles sont donc des entrées pour HERBS. Malgré tout, le système est suffisamment flexible pour permettre l'intégration d'algorithmes d'extraction de règles. Entre temps, différentes plate-formes ont vu le jour, par exemple ARQAT [Huynh et al., 2006]. Pour disposer de résultats en situation, il est nécessaire de pouvoir évaluer des bases de règles sur des bases de cas, puis d'en évaluer l'intérêt au moyen de mesures, afin d'analyser le comportement de ces mesures en situation, et de les comparer.

Ces constats ont donc motivé le développement de HERBS. Celui-ci a pour but de permettre d'effectuer des études en procédant de manière itérative via une interface graphique, ou de systématiser des campagnes expérimentales par des appels en ligne de commande. Une attention particulière a été portée lors de sa conception sur des questions de performances, et de portabilité.

1 L'outil de simulation HERBS

Un premier prototype [Vaillant et al., 2002, Vaillant et al., 2003] a mené à la définition d'un ensemble de fonctionnalités, et a été récemment ré-implémenté intégralement[1] [Vaillant et al., 2006b]. La version actuelle dispose de deux méthodes d'accès, dédiées à des traitements spécifiques : la première permet d'effectuer une exploration pas à pas, la seconde

[1] Nous tenons à remercier Sébastien Bigaret pour sa contribution apportée à la nouvelle version de HERBS. Sans son aide, elle n'aurait certainement pas vu le jour. Cette version a été présentée lors de la session "atelier logiciel" d'EGC 2005.

permet de mettre en œuvre des campagnes expérimentales. Les résultats intermédiaires sont sauvegardés, afin de limiter le recalcul de grandeurs déjà évaluées.

Mettre en place des campagnes expérimentales sur des volumes conséquents mène très directement à se poser des questions d'optimisation, au vu des nombreux calculs à effectuer. Certains calculs, comme l'évaluation des règles sont clairement redondants : si l'on souhaite croiser les valeurs prises par plusieurs mesures sur une même base de règle, étant donné une base de données, il peut être intéressant de disposer d'un mécanisme de cache, pour mémoriser les grandeurs caractéristiques des règles, et ne pas les recalculer pour chacune des mesures.

Ainsi, pour fournir une telle étendue de services, HERBS s'appuie sur un SGBDR, est pourvu d'une interface graphique conviviale, mais peut également être utilisé en ligne de commande. Le SGBDR permet le stockage des données, ainsi que d'évaluer efficacement des règles sur des bases de cas, au moyen de requêtes SQL, dans un environnement ensembliste. Le cœur du système permet l'exécution de traitements atomiques de tâches, et assure la modularité de l'outil. Ce cœur est alors interrogé au besoin, soit par l'interface graphique, soit par la commande en ligne. La figure IV.1 présente l'architecture de HERBS.

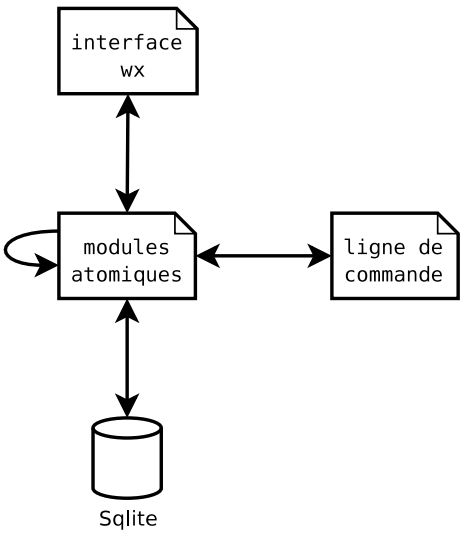


FIG. IV.1 – Architecture de l'outil HERBS

Le choix des langages a été guidé par deux aspects majeurs précédemment mentionnés, la portabilité, et l'intégration.

Le gestionnaire de base de données est SQLITE (`www.sqlite.org`), gestionnaire portable et épuré de nombreuses fonctionnalités dont disposent la plus part des autres SGBDR, inutiles dans notre cas. L'un des atouts majeurs, est qu'il ne nécessite aucune configuration, ce qui semble très intéressant pour de futures utilisations, par des utilisateurs non experts en configuration de clients/serveurs de bases de données. Il a de plus été modifié afin d'intégrer des calculs spécifiques à notre problématique (voir section 1.2). Cette version modifiée fonctionne et a été testée sous linux et windows.

Le code est majoritairement écrit en python (`www.python.org`), certaines parties sont

en C pour des raisons de performances, et l'interface fait appel à la librairie wxWidgets (`www.wxwidgets.org`). Ces composants sont disponibles sur de très nombreuses plate-formes et environnements graphiques.

1.1 Importation de données

HERBS fonctionnant en post-analyse, il est nécessaire de lui fournir en entrée les données sur lesquelles on souhaite travailler. Ces données sont les bases de cas et de règles, qui sont regroupées en couples compatibles (traitant des mêmes attributs) automatiquement.

Afin d'importer des données, on créée ou on réutilise un projet existant. Ce projet consiste en un fichier, qui est la base de données SQLITE qui sera utilisée. Les figures IV.2 et IV.3 illustrent l'import de données d'un couple de base de cas et de règles compatibles. Dans le mode graphique, il est actuellement impossible de préciser le nom et le type des attributs de la base de cas, les noms doivent être mis sur la première ligne de la base de cas, et leur type est déterminé automatiquement. Il est toutefois possible de le faire via la ligne de commande, comme le montre le script shell IV.1. Nous importons ici la base de cas "car", disponible sur l'UCI Repository [Newman et al., 1998], à laquelle nous avons appliqué l'algorithme APRIORI afin de générer la base de règles. Avec des paramètres de support et de confiance minimaux de 5% et 60% respectivement, nous avons obtenu 145 règles. Nous illustrerons l'utilisation de HERBS avec ce couple de base de cas et de règles.

Listing IV.1 – Création d'un projet HERBS en ligne de commande

```
db="car.hrb"
# Création de la base SQLITE
herbscmd.py -c $db
# Importation de la base de cas en précisant le nom
# des attributs et leur type
herbscmd.py $db dataset import --csv -n car \
    -l buying:maint:doors:persons:lug_boot:safety:class \
    -t text:text:text:text:text:text:text car.csv
# Importation de la base de règles
herbscmd.py $db rulesset import -a -n car car.rules
```

1.2 Traitements mono-mesure

Une fois un couple base de cas/base de règles importé dans le projet, il est possible d'évaluer les règles sur la base de cas, puis de calculer les valeurs prises par les mesures. Les couples pour lesquelles ces évaluations sont effectuées sont marqués comme évalués dans l'interface graphique, le menu contextuel (figure IV.6) est alors modifié en conséquence, rendant disponible le calcul de mesures (figure IV.4). Les évaluations pouvant être longues, une barre d'avancement (figure IV.5) permet d'estimer le temps de calcul restant.

FIG. IV.2 – Importation d'une base de cas

FIG. IV.3 – Importation d'une base de règle

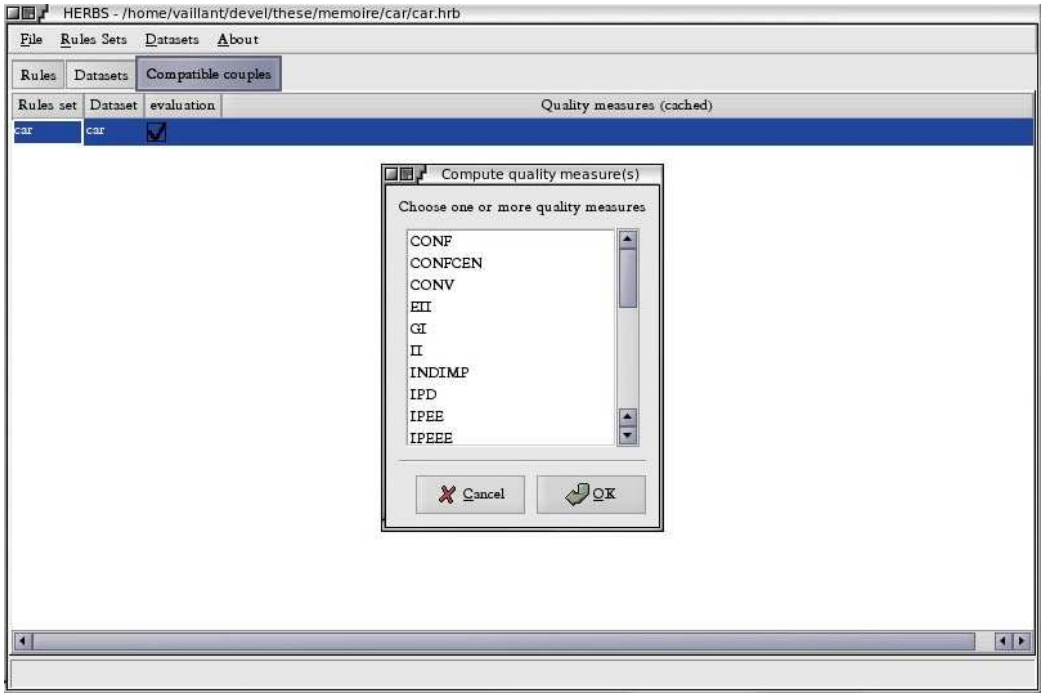


FIG. IV.4 – Choix des mesures à évaluer

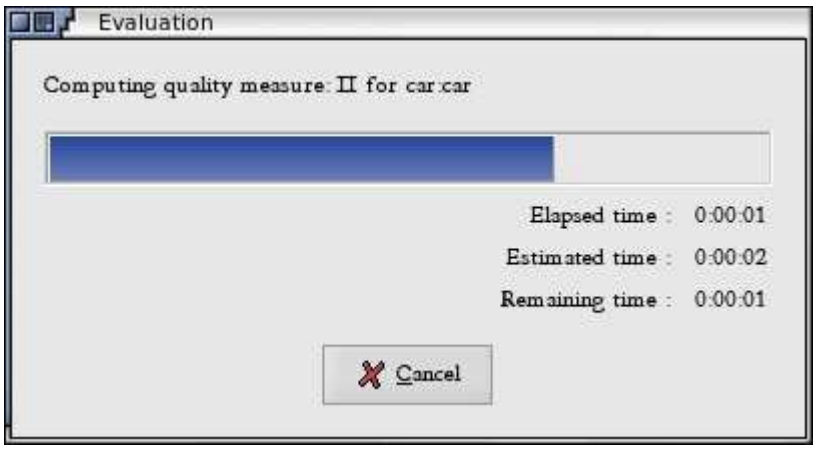


FIG. IV.5 – Barre d'avancement des calculs

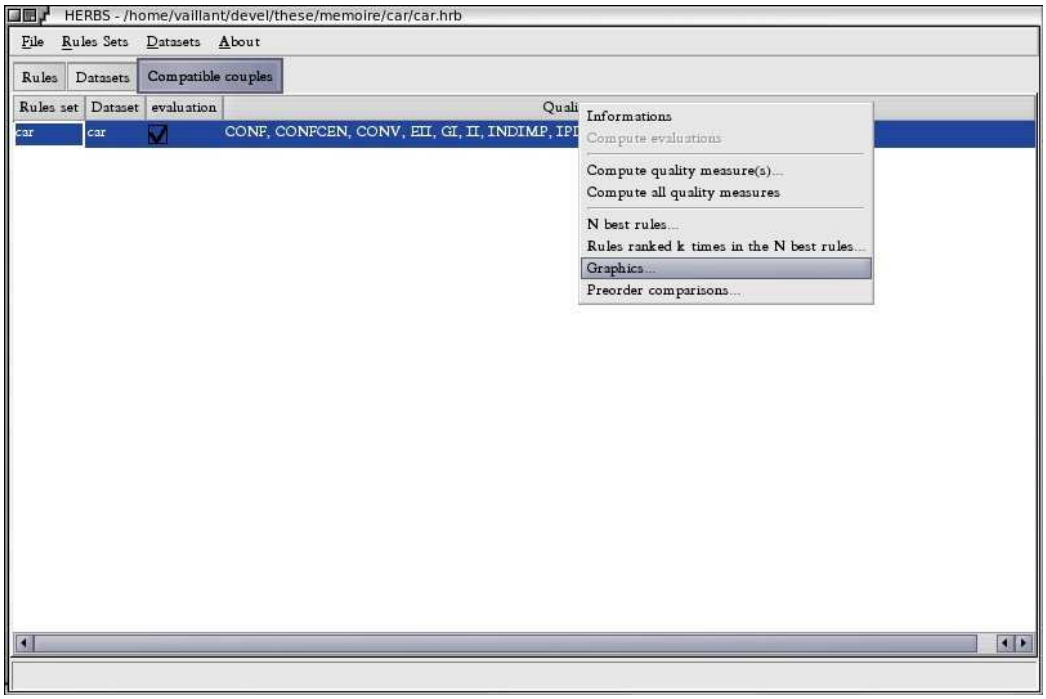


FIG. IV.6 – Menu contextuel associé à un couple de base de cas et de règles

1.2.1 Sélection des N meilleures règles

Le but de la "sélection des N meilleures règles" est de retourner, étant donné une mesure de qualité et un couple de base de cas et de base de règles, l'ensemble des règles les mieux classées par la mesure, la taille minimale de cet ensemble étant paramétrée par l'entier N.

Pour N fixé, on range les règles en fonction de la valeur de la mesure de qualité, et on retourne un sous-ensemble de règles tel que le rang de chacune des règles du sous-ensemble soit inférieur ou égal à N (par rang on entend le numéro de rangement de la règle : la mieux classée est de rang 1, la seconde de rang 2,..., avec d'éventuels ex-aequo).

L'obtention de cet ensemble peut permettre de générer de nouvelles bases de règles, ou d'approcher de manière relativement naïve le potentiel discriminant d'une mesure. Une mesure non-discriminante aura tendance à regrouper les règles en "paquets", alors qu'une mesure discriminante s'étalera mieux sur la plage de valeurs qu'elle peut prendre. Ainsi, en observant l'évolution de la taille effective de l'ensemble des N meilleures solutions en fonction de N, on dispose en première approche d'une évaluation du caractère discriminant ou non d'une mesure.

La figure IV.7 présente l'évolution de la taille de cet ensemble en fonction de N pour les mesures CONF, MOCO et TIIE. Si l'évolution de la taille de cet ensemble semble assez similaire pour CONF et MOCO dès que N dépasse 60, on distingue clairement une nette différence de comportement de TIIE lorsque N est grand. Cette différence correspond aux règles écartées par l'indice d'inclusion (*i.e.* évaluées à 0).

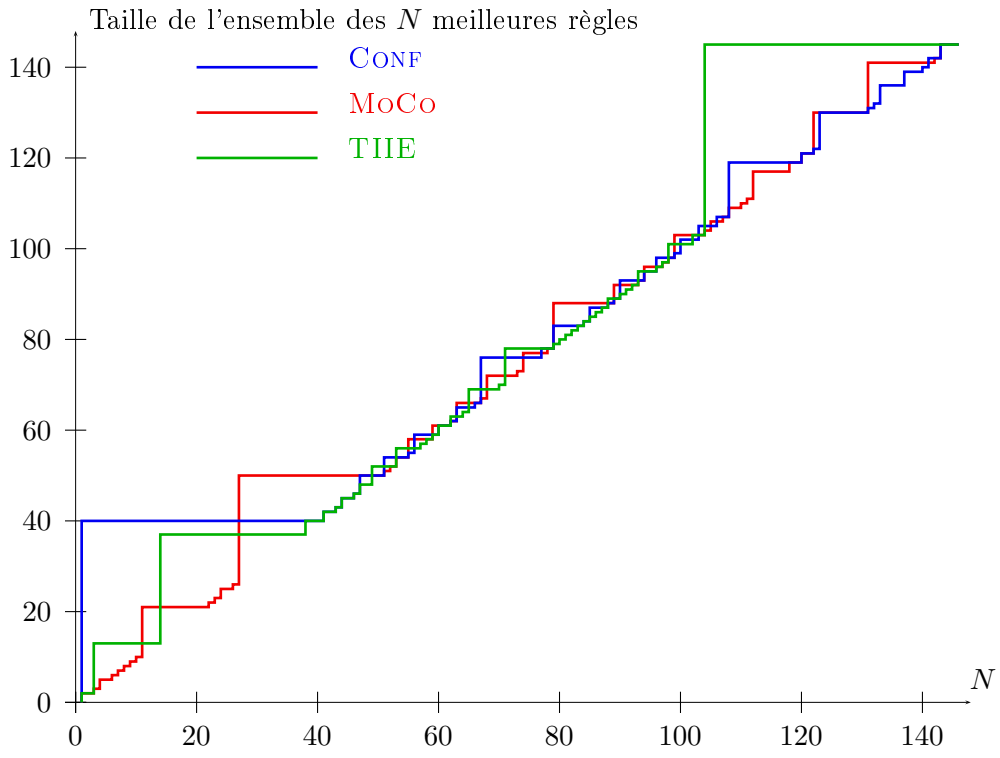


FIG. IV.7 – Evolution de l'ensemble des N meilleures règles pour CONF, MOCO et TIIE

1.2.2 Distribution des valeurs d'une mesure

L'analyse d'une mesure de qualité peut se faire en étudiant les valeurs prises par cette mesure sur un couple de base de cas et de base de règles donné. On souhaite dans ce cas tracer les valeurs prises par la mesure pour chacune des règles de la base de règle, en ordonnée. Afin d'obtenir une représentation en deux dimensions, on peut utiliser diverses grandeurs en abscise.

Listing IV.2 – Tracé de mesures de qualité en fonction de $n_{a\bar{b}}$

```
db="car.hrb"
# Tracé de différentes mesures en fonction de nab_
for mesure in CONF CONFCEN CONV EII IPD INDIMP IPEE SUP
do
   herbscmd.py $db dist 1 -r car -d car \
     -x nab_ -y $mesure -o $mesure.eps
done
```

Le menu contextuel `graphique...` permet de visualiser la distribution d'une mesure de qualité selon les grandeurs n_a, n_b, n_{ab} ou $n_{a\bar{b}}$. Il est possible d'effectuer des zooms sur des zones particulières, et de se déplacer dans le plan. En mode ligne de commande, on peut préciser un fichier dans lequel enregistrer le graphique obtenu, ou lancer l'affichage graphique au moyen de l'option `show`. Le script IV.2 génère les neuf images de la figure IV.8. Bien que

les dimensions de nos bases de cas et de règles soient petites, on retrouve pour certaines mesures les tracés théoriques du chapitre II, par exemple pour IPEE et IIE.

Au delà des tracés en situation, HERBS permet également d'effectuer les tracés théoriques de certaines mesures. Ainsi, les différentes courbes représentatives des mesures probabilistes ont été obtenues en utilisant la version modifiée de SQLITE. Des soucis d'optimisation de temps de calcul ont mené à intégrer le calcul de ces mesures en tant que fonctions SQL. On peut donc tout à fait interroger SQLITE ainsi (la fonction SQLITE `imp_idx` calcule la valeur de l'intensité d'implication, INTIMP) :

```
sqlite> select imp_idx(2000, 500, 800, 220);
0.87574991553671
```

Ainsi, pour obtenir les valeurs de l'intensité d'implication, pour $n_a = 500$, $n_b = 800$, $n = 2000$, en fonction de $n_{a\bar{b}}$, il suffit d'exécuter le script IV.3, qui retourne la liste de points $(p_{a\bar{b}}, \text{INTIMP}(n, n_a, n_b, n_{ab}))$ (voir extraits table IV.1).

($p_{a\bar{b}}$, INTIMP)
(0, 1)
(0.005, 1)
(0.01, 1)
...
(0.08, 1)
(0.085, 0.999999999999969)
(0.09, 0.999999999997866)
(0.095, 0.999999999892343)
...
(0.13, 0.989528209378679)
(0.135, 0.958363218779878)
(0.14, 0.87574991553671)
(0.145, 0.718030272238959)
(0.15, 0.5)
(0.155, 0.281969727761041)
(0.16, 0.12425008446329)
(0.165, 0.0416367812201217)
(0.17, 0.0104717906213208)
(0.175, 0.00195086078626072)
(0.18, 0.000266103255337602)
...
(0.225, 0)
(0.23, 0)
...
(0.265, 0)

TAB. IV.1 – Valeurs de INTIMP en fonction de $p_{a\bar{b}}$, avec $n_a = 0.25$ et $n_b = 0.4$

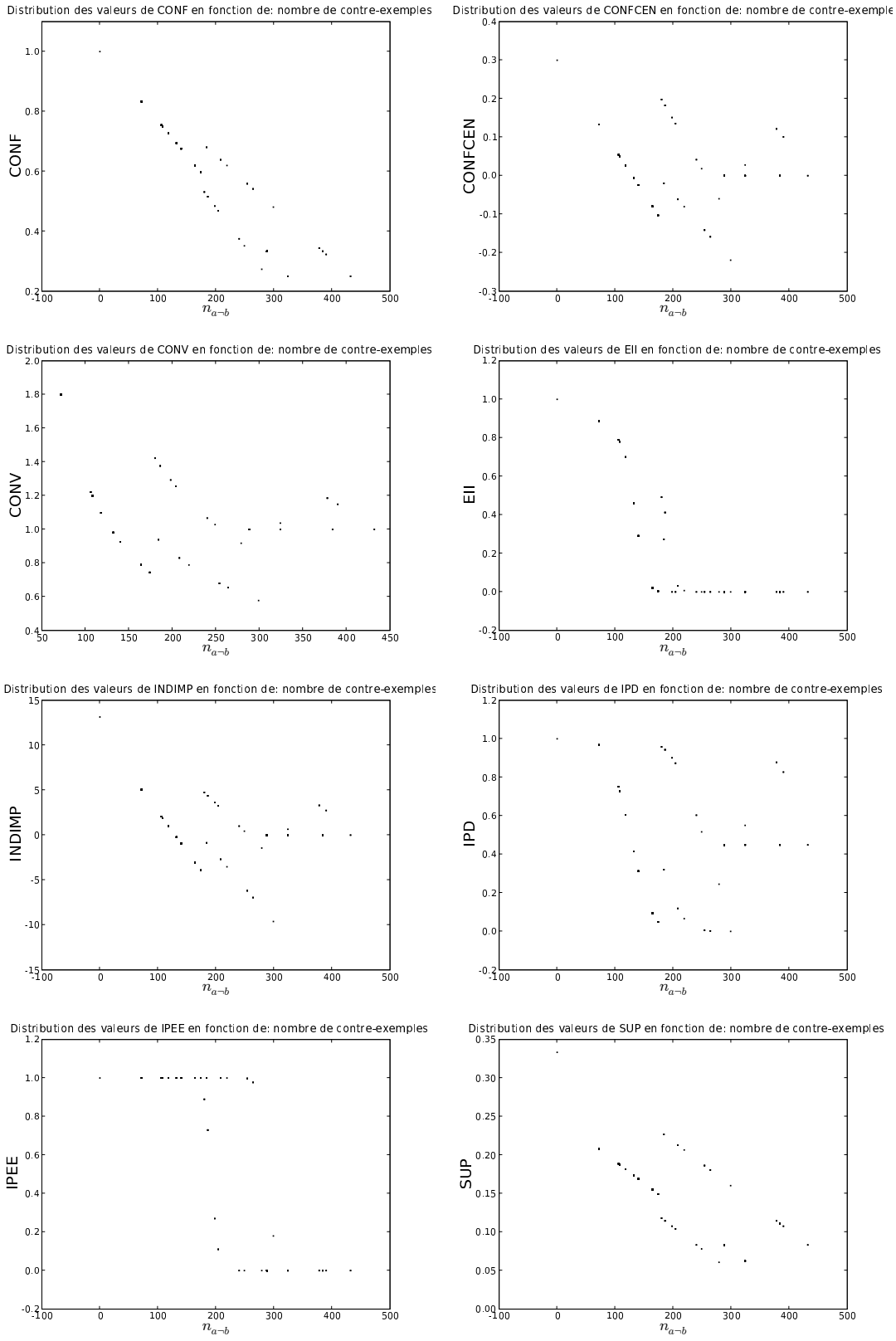


FIG. IV.8 – Variations de 8 mesures en fonction de $n_{a\bar{b}}$

Listing IV.3 – Evolution de IntImp en fonction de $n_{a\bar{b}}$

```
na=500; nb=800; n=2000
# valeurs de IntImp en fonction de nabb (contre-exemples)
for ((nabb=0; nabb<540; nabb+=10))
do
  echo \
   "select 1.0*$nabb/$n, imp_idx($n, $na, $nb, $na - $nabb);"
done |sqlite |awk 'BEGIN {FS = "|"} {print "(" $1 "," $2 ")";}'
```

1.3 Traitements multi-mesures

Au delà de résultats permettant d'analyser le comportement d'une mesure sur un couple de bases de cas et de règles, nous nous sommes intéressés à des traitements prenant en compte plusieurs mesures, afin de croiser leurs comportements.

1.3.1 Distribution croisée des valeurs de deux mesures

Les précédents résultats graphiques présentant les variations d'une mesure en fonction des grandeurs caractéristiques ne mettent pas toujours suffisemment en évidence le comportement concordant ou discordant de deux mesures. Ainsi, il nous a semblé nécessaire de proposer à l'utilisateur de visualiser sur un couple de bases de cas et de règles donné la distribution croisée des valeurs prises par deux mesures.

En effet, nos mesures – toutes décroissantes avec $n_{a\bar{b}}$ – revètent des allures "similaires" si on les observe en fonction de cette grandeur, au sens où plus $n_{a\bar{b}}$ sera grand, plus la valeur prise par la mesure sera faible. Pour autant, il est fréquent que deux mesures n'ordonnent pas l'ensemble des règles de la même manière. La représentation croisée des valeurs qu'elles prennent permet de mieux mettre en évidence leurs comportements, semblable ou non.

De manière tout à fait similaire au tracé des distributions d'une mesure, on peut utiliser Herbs en mode interactif pour procéder à des comparaisons pas à pas entre couples de mesures, ou en mode commande pour réaliser des campagnes expérimentales.

Reprenant les trois mesures Conf, MoCo et TIIE, le script IV.4 permet de croiser les distributions des valeurs qu'elles prennent sur la base de règle car. Les résultats sont présentés en figures IV.9 à IV.11.

Ces tracés nous permettent d'affiner la connaissance que nous avions de ces mesures via l'analyse de l'évolution des ensembles des N meilleures règles. MoCo et Conf (figure IV.9), pour lesquelles la taille de cet ensemble évoluait de manière semblable ne rangent pas les règles dans le même ordre. En revanche, la courbe croisant TIIE et Conf (figure IV.10) indique que les règles bien classées par l'une le sont également par l'autre, les différences apparaissant dans le bas du classement. On voit à nouveau l'effet de l'évaluation à 0 des règles ne passant pas soit l'indépendance, soit l'indétermination. Il est manifeste que Conf conserve certaines de ces règles pour un seuil de 60%. La même correction est visible lorsqu'on compare TIIE à MoCo (figure IV.11). Ce comportement nous semble surprenant, ces mesures étant si différentes d'un point de vue formel. Nous l'expliquons ici par le faible

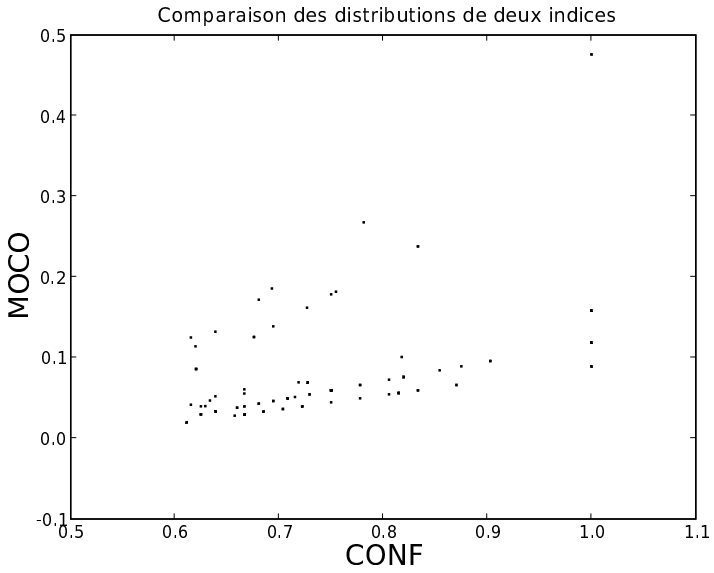


FIG. IV.9 – Variations croisées de CONF et MOCO

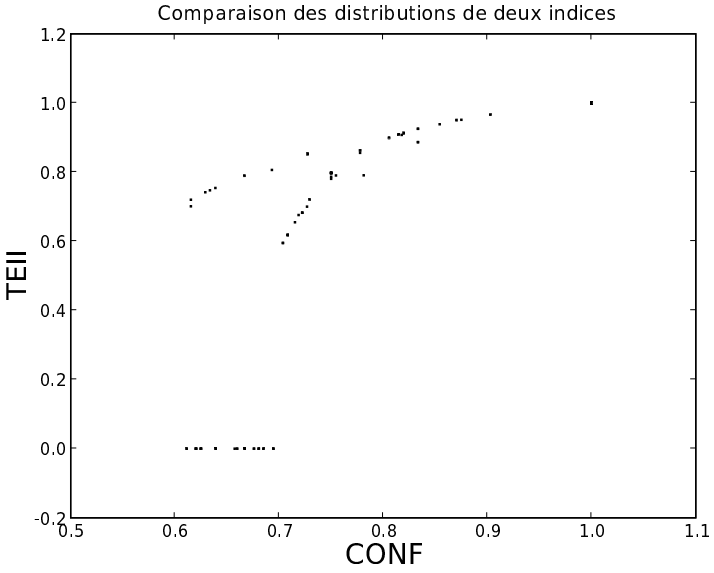


FIG. IV.10 – Variations croisées de CONF et TIIE

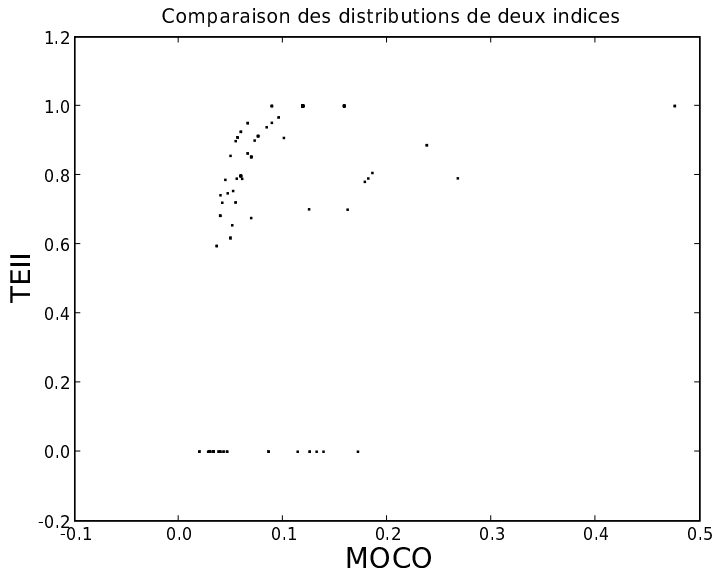


FIG. IV.11 – Variations croisées de MoCo et TIIE

nombre de règles, et des expériences sur des bases plus conséquentes ont montré des comportements en situation plus divergents.

Listing IV.4 – Evolutions comparées de CONF, MoCo et TIIE

```
MESURES=( CONF MOCO TEII )
db="car.hrb"
for ((i=1; i<=3; i++))
do
    for ((j=i+1; j<=3; j++))
    do
        herbscmd.py $db dist 1 -r car -d car -x $MESURES[$i] \
           -y $MESURES[$j] -o $MESURES[$i]-$MESURES[$j].eps
    done
done
```

1.3.2 Comparaison de préordres induits par deux mesures

Afin de disposer de résultats numériques, au delà des distributions croisées, pour mesurer la force de l'accord entre deux mesures de qualité sur l'ordonnancement des règles, nous utilisons une fonction particulière, un coefficient d'accord entre préordres.

Les mesures de qualité induisent naturellement des préordres totaux sur les règles évaluées, en rangeant ces dernières de la mieux évaluée à la moins bien évaluée. On peut alors calculer un coefficient d'accord. Il n'est pas sensible aux écarts de valeurs prises par les mesures, mais résume la force de l'accord sur l'ordre qu'elles imposent : son calcul se base sur

l'analyse des rangs.

Dans [Giakoumakis and Monjardet, 1987], les auteurs rappellent la définition de seize coefficients d'accord entre préordres totaux, dérivés soit de dissimilarités, soit de produits scalaires. Ces coefficients sont des généralisations du τ de Kendall ou du ρ de Spearman. Nous n'utiliserons ici qu'un seul de ces coefficients, τ_1, qui nous semble le mieux adapté à notre approche de par ses propriétés, que nous allons expliciter ci-dessous. Les autres coefficients sont définis sur les mêmes grandeurs, mais ne vérifient pas les mêmes propriétés.

Un préordre total $\mathcal{P}$ défini sur un ensemble X se décompose en deux composantes, $\mathcal{I}$ (la partie symétrique du préordre total), et $\mathcal{O}$ (la partie asymétrique du préordre total) :

$$\mathcal{P} = \mathcal{I} + \mathcal{O}$$

avec :

$$\mathcal{I} = \{(x, y) \in X^2 \text{ tel que } x\mathcal{P}y \text{ et } y\mathcal{P}x\} \text{ (partie symétrique de } \mathcal{P})$$

et

$$\mathcal{O} = \{(x, y) \in X^2 \text{ tel que } x\mathcal{P}y \text{ et } \neg y\mathcal{P}x\} \text{ (partie asymétrique de } \mathcal{P})$$

$\mathcal{P}$ induit une partition de l'ensemble des couples de X :

$$X^2 = \mathcal{O} + \mathcal{I} + \mathcal{O}^r$$

où $\mathcal{O}^r$ est la relation réciproque de $\mathcal{O}$:

$$(x, y) \in \mathcal{O}^r \Leftrightarrow (y, x) \in \mathcal{O}$$

Etant donné la paire $(\mathcal{P}, \mathcal{P}')$ de préordres totaux sur X, et un couple (x, y) de X^2, on dira que :

- (x, y) est un *accord strict* si et seulement si $(x, y) \in \mathcal{O} \cap \mathcal{O}'$,
- (x, y) est un *accord large* si et seulement si $(x, y) \in \mathcal{I} \cap \mathcal{I}'$,
- (x, y) est un *semi-accord* si et seulement si $(x, y) \in (\mathcal{O} \cap \mathcal{I}') \cup (\mathcal{I} \cap \mathcal{O}')$,
- (x, y) est un *semi-désaccord* si et seulement si $(x, y) \in (\mathcal{I} \cap \mathcal{O}'^r) \cup (\mathcal{O}^r \cap \mathcal{I}')$,
- (x, y) est un *désaccord strict* si et seulement si $(x, y) \in (\mathcal{O} \cap \mathcal{O}'^r) \cup (\mathcal{O}^r \cap \mathcal{O}')$,
- (x, y) est un *co-accord* si et seulement si $(x, y) \in \mathcal{O}^r \cap \mathcal{O}'^r$.

On peut donc comparer les deux préordres $\mathcal{P}$ et $\mathcal{P}'$ selon ces cinq paramètres, repris en table IV.2 :

- a est le nombre d'accords stricts (a est aussi le nombre de co-accord).
- c est le nombre d'accords larges.
- $2d$ est le nombre de désaccords stricts.
- $b = b_1 + b_2$ est le nombre de semi-accords (b est aussi le nombre de semi-désaccords).

On pose de plus $c^* = \frac{c-n}{2}$ (avec $n = |X|$). Ce recentrage est intéressant puisque pour tout élément x de X, (x, x) est nécessairement un accord large, et si $(x, y) \in X^2$ est un accord large, alors (y, x) l'est également.

Pour évaluer les paramètres a, b, c et d, on doit considérer les éléments de X^2. Etant donné la symétrie du calcul, on peut se limiter à l'étude de $n(n-1)/2$ couples. La complexité d'une telle évaluation se fait en $\mathcal{O}(n^2)$.

Avec ces notations, le coefficient d'accord τ_1 est défini comme :

$$\tau_1 = \frac{2(a-d) + 2(c^* - b)}{n(n-1)}$$

Il prend ainsi ses valeurs dans $[-1, 1]$. De plus, il vérifie des propriétés que nous souhaitions. Il est symétrique selon $\mathcal{P}$ et $\mathcal{P}'$. Sa valeur est maximale si et seulement si $\mathcal{P} = \mathcal{P}'$ (il n'y a dans ce cas que des accords larges ou stricts). Sa valeur est minimale si et seulement si $\mathcal{P} \cap \mathcal{P}'^r$ est un ordre total.

L'utilisation de cette méthode de comparaison est illustrée en section 2.

	$\mathcal{O}'$	$\mathcal{I}'$	$\mathcal{O}'^r$
$\mathcal{O}$	a	b_2	d
$\mathcal{I}$	b_1	c	b_1
$\mathcal{O}^r$	d	b_2	a

TAB. IV.2 – Répartition des couples d'éléments de X^2 selon deux préordres $\mathcal{P}$ et $\mathcal{P}'$

1.3.3 Règles sélectionnées k fois parmi les meilleures

La dernière voie que nous avons explorée en vue de filtrer une base de règles met en œuvre un système de "vote".

Etant donné un ensemble de m mesures, et deux entiers k et N, on peut souhaiter savoir quelles sont les règles qui sont sélectionnées au moins k fois par les m mesures comme étant de rang au plus N.

Cette possibilité étend le précédent filtrage des N meilleures règles à un cas généralisé à m mesures.

Si l'on prend $k = 1$, on obtient alors l'union des ensembles des N meilleures règles pour les m mesures, puisque toutes les règles sont sélectionnées au moins une fois. Pour $k = m$, on obtient au contraire leur intersection.

Ainsi, nous disposons d'un moyen paramétrique simple pour agréger plus ou moins strictement les avis de plusieurs mesures concernant l'appartenance ou non d'une règle à l'ensemble de celles étant d'intérêt.

Tout comme dans le cas de la comparaison de préordres, il est à noter que cette démarche se démarque fondamentalement de l'approche théorique présentée au chapitre III, au sens

où nous ne tenons pas compte ici de la valeur que prend la mesure, mais uniquement de l'ordre dans lequel cette dernière range les règles.

A nouveau, nous allons illustrer ceci au moyen des mesures Conf, MoCo et TIIE appliquées à la base car. Nous choisissons d'observer l'évolution de l'ensemble des règles sélectionnées k fois parmi les meilleures en fonction de k.

La figure IV.12, dont la liste des points nécessaires pour son tracé est fournie par le script IV.5, présente l'évolution des tailles de ces ensembles de règles, pour k allant de 1 à 3.

Il est à noter que certaines règles (il s'agit ici de règles logiques) sont rapidement intégrées au sous-ensemble de règles sélectionné trois fois par les mesures (dès $N = 27$). Il faut ensuite aller jusqu'à $N = 53$ (soit environ un tiers du nombre total de règles) pour que les suivantes soient intégrées à cet ensemble.

Listing IV.5 – Nombre de N meilleures règles sélectionnées k fois par Conf, MoCo et TIIE

```
db="car.hrb"
for ((k=1; k<=3; k++))
do
    old_nb=0
    old_N=1
    for ((N=1; N<145; N++))
    do
        nb=‘herbscmd.py $db best 1 -r car -d car \
            -m CONF:MOCO:TEII -N $N -k $k | wc -l‘
        nb=‘expr $nb - 1‘
        if [ $nb != $old_nb ]
        then
           echo -n "($N,$old_nb)"
           echo -n "($N,$nb)"
           old_nb=$nb
           old_N=$N
        fi
    done
    echo "($N,$N)"
done
```

2 Expérimentations sur des bases de données réelles

Afin d'étudier en pratique le comportement de mesures de qualité, nous les avons confronté à des données réelles, issues de l'UCI Repository [Newman et al., 1998].

Nous avons utilisé Apriori afin de générer les bases de règles associées à ces bases de données. Nous présentons ci-dessous les données, ainsi que les résultats obtenus. Ces résultats expérimentaux permettent de proposer une classification expérimentale des mesures de qualité, qu'il est intéressant de confronter à celle obtenue sur des critères formels [Vaillant

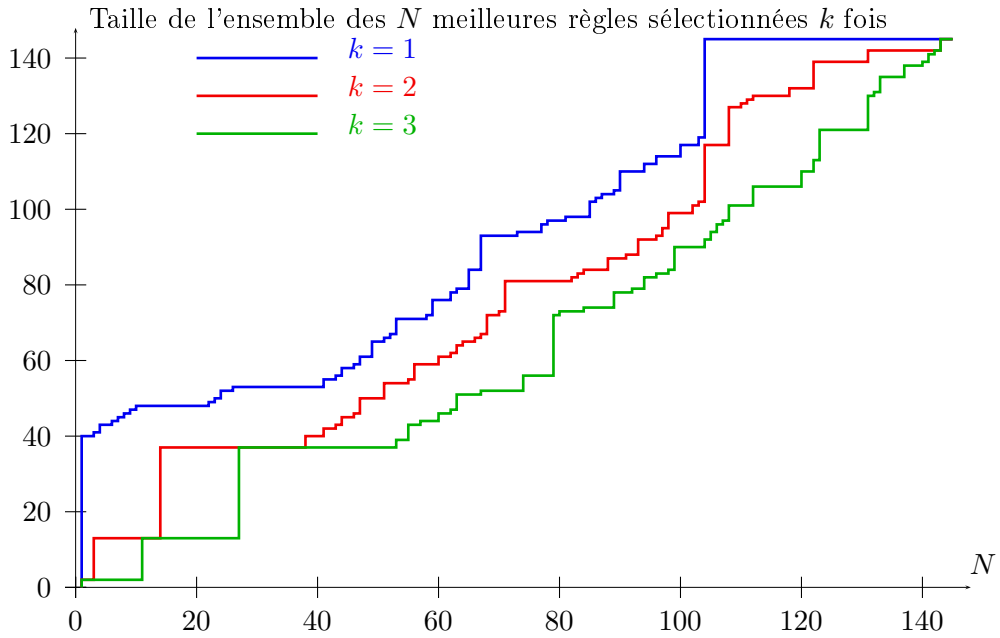


FIG. IV.12 – Evolution de l'ensemble des N meilleures règles sélectionnées k fois

et al., 2004].

2.1 Présentation des bases de données retenues

On rappelle que nous nous situons en post-analyse, et que les bases de cas et de règles sont donc des entrées. Le problème de la qualité des données n'a pas été abordé, et nous considérons qu'il s'agit d'un travail fait en amont lors du processus de fouille de données. Le lecteur intéressé par la qualité des données et les métriques associées pourra se référer à [Berti-Équille, 2004] par exemple.

Chaque base choisie a été transformée afin de l'adapter à l'algorithme APRIORI d'une part, et à l'entrée de HERBS d'autre part. Ainsi, pour un fichier de données contenant des valeurs $\texttt{att}_{i,j}$ associées aux attributs $\texttt{att}_i$, comme présenté en table IV.3, nous précisons à quel attribut se réfère les valeurs, tel que présenté en table IV.4.

Ce recodage est nécessaire tout d'abord parce qu'APRIORI prend en considération des items, et non des modalités d'attributs. Si deux attributs partagent une ou plusieurs modalités (ce qui est le cas dans les bases que nous considérons), APRIORI considérera ces modalités comme faisant référence à la même occurrence de donnée présente, sans distinguer l'attribut auquel il se rattache. Le recodage assure donc la transformation de nos données initialement sous forme tabulaire en un ensemble de co-occurrences d'items. Le fait que nous restions sous une forme tabulaire de mêmes dimensions tient du fait que nous avons procédé à un encodage binaire disjonctif complet des modalités prises par chaque attribut à partir de données déjà sous forme tabulaire, en omettant les absences.

Par ailleurs, ce codage nous est nécessaire lors de l'importation de règles dans HERBS, afin de détecter quelles modalités réfèrent à quel attribut, et quels attributs sont considérés

par les règles (c'est cette analyse qui nous permet de détecter automatiquement les couples de bases de cas et de règles compatibles).

Les données reformulées ainsi, APRIORI générera des règles de la forme :
$\mathtt{att}_2 = \mathtt{att}_{2,3}$ <- $\mathtt{att}_1 = \mathtt{att}_{1,2}$

$\mathtt{att}_{1,2}$	$\mathtt{att}_{2,1}$	$\mathtt{att}_{3,5}$
$\mathtt{att}_{1,2}$	$\mathtt{att}_{2,3}$	$\mathtt{att}_{3,4}$
$\mathtt{att}_{1,1}$	$\mathtt{att}_{2,4}$	$\mathtt{att}_{3,3}$
$\mathtt{att}_{1,2}$	$\mathtt{att}_{2,2}$	$\mathtt{att}_{3,2}$
$\mathtt{att}_{1,3}$	$\mathtt{att}_{2,2}$	$\mathtt{att}_{3,1}$

TAB. IV.3 – Présentation des données brutes

$\mathtt{att}_1 = \mathtt{att}_{1,2}$	$\mathtt{att}_2 = \mathtt{att}_{2,1}$	$\mathtt{att}_3 = \mathtt{att}_{3,5}$
$\mathtt{att}_1 = \mathtt{att}_{1,2}$	$\mathtt{att}_2 = \mathtt{att}_{2,3}$	$\mathtt{att}_3 = \mathtt{att}_{3,4}$
$\mathtt{att}_1 = \mathtt{att}_{1,1}$	$\mathtt{att}_2 = \mathtt{att}_{2,4}$	$\mathtt{att}_3 = \mathtt{att}_{3,3}$
$\mathtt{att}_1 = \mathtt{att}_{1,2}$	$\mathtt{att}_2 = \mathtt{att}_{2,2}$	$\mathtt{att}_3 = \mathtt{att}_{3,2}$
$\mathtt{att}_1 = \mathtt{att}_{1,3}$	$\mathtt{att}_2 = \mathtt{att}_{2,2}$	$\mathtt{att}_3 = \mathtt{att}_{3,1}$

TAB. IV.4 – Présentation des données recodées

Nous avons retenu pour procéder à notre étude en situation les bases de cas suivantes : *Auto-mpg*, *Breast Cancer Winsconsin*, *Car*, *Contraceptive Method Choice*, et *Solarflare*. Certaines de ces bases de cas contenant des données manquantes, les enregistrements concernés ont simplement été supprimés. La base *Solarflare* disponible sur le site de l'UCI est scindée en deux, et nous ferons donc référence à *Solarflare1* et *Solarflare2* selon la base de cas considérée.

La table IV.5 liste le nombre d'enregistrements et d'attributs (hors identifiants de transactions) pour chacune des bases, ainsi que l'abréviation que nous utiliserons par la suite pour y faire référence pour celles ayant de longs noms.

Base de cas	nombre de cas	nombre d'attributs	nom abrégé
Auto-mpg	392	8	*Auto-mpg*
Breast Cancer Winsconsin	683	10	*Bcw*
Car	1728	7	*Car*
Contraceptive Method Choice	1473	10	*Cmc*
Solarflare1	323	13	SF_1
Solarflare2	1066	13	SF_2

TAB. IV.5 – Présentation des données recodées

Nous avons utilisé APRIORI afin de générer des bases de règles associées à ces bases de cas, avec les paramètres listés en table IV.6.

Afin d'observer l'impact d'opérations de filtrage sur les bases de règles nous avons effectué deux filtrages [Vaillant et al., 2004].

Base de cas	seuil de support	seuil de confiance	nombre de règles
Auto-mpg	5%	50%	49
Bcw	10%	70%	3095
Car	5%	60%	145
Cmc	5%	60%	2878
SF_1	20%	85%	5402
SF_2	20%	85%	3346

TAB. IV.6 – Présentation des bases de règles extraites par APRIORI

Le premier filtrage que nous avons effectué ne retient que les règles ayant un LIFT supérieur à 1.2. Nous l'avons appliqué à la base *Cmc*, et obtenons ainsi une deuxième base de règle *Cmc2*. Celle-ci contient 766 règles. Le script IV.6 nous permet d'effectuer cette opération.

Listing IV.6 – Filtrage de *Cmc* selon LIFT

```
db="car.hrb"
# on crée un fichier temporaire dont la structure
# est identique à celle de ceux créés par apriori
herbscmd.py $db comp -S LIFT:cmc:cmc | \
  awk 'BEGIN {FS="[(,)]";} \
       {if ($2>1.2) print $4 " <- " $3 " (0\%, 0\%)";}' | \
  sed -e 's/ and / /g' > tmp.rules
# on importe les règles
herbscmd.py $db rulesset import -a -n cmc2 tmp.rules
# le fichier temporaire peut être effacé
rm tmp.rules
```

Pour le second filtrage, nous souhaitons retenir les règles qui renforcent ce que l'on sait de `B` *a priori*, c'est-à-dire les règles telles que $p_{\mathtt{B}|\mathtt{A}} > p_{\mathtt{B}}$. Nous insistons sur cette propriété des règles en filtrant non pas à l'indépendance, mais selon une modélisation de la loi de n_{ab}, comme proposé dans [Lallich et al., 2004].

Selon le modèle hypergéométrique (voir dans le tableau récapitulatif II.7 reprenant la modélisation $\mathcal{N}_1$, en section 7), la variable aléatoire $\mathcal{N}_{ab}$ suit la loi $H(n, n_a, p_b)$. Dans les conditions standard d'approximations (*i.e.* $\frac{n_a n_b}{n} \geq 5$ et $\frac{n_a n_{\bar{b}}}{n} \geq 5$), on montre qu'alors :

$$P(\mathcal{N}_{ab} \geq n_{ab}) = 1 - \Phi(\sqrt{(n)}\mathrm{R})$$

où Φ est la fonction de répartition de la loi normale centrée réduite.

En calculant le coefficient de corrélation associé à `A` et `B`, on peut donc obtenir la p-valeur de $n_{a\bar{b}}$. Prenant un risque à 5%, on ne retiendra alors que les règles telles que :

$$\mathrm{R} \geq \frac{1.645}{\sqrt{n}}$$

Nous avons appliqué ce filtrage sur la base de règles issue de SF_1. Par soucis de clareté, nous introduisons les notations suivantes :

- R_1 est la base de règle issue de SF_1,
- R_2 est la base de règle issue de SF_2,
- R_1^1 est la base de règle issue de SF_1 et filtrée sur SF_1,
- R_1^2 est la base de règle issue de SF_1 et filtrée sur SF_2.

R_1^1 est composée de 2152 règles, et R_1^2 de 2005 règles.

2.2 Présentation de résultats obtenus avec HERBS

Nous avons calculé les coefficients d'accords entre préordres pour les vingt mesures retenues dans notre étude formelle (voir chapitre III) et dix couples de bases de cas et de règles. Ces résultats sont synthétisés en table IV.7. Il convient de lire ce tableau comme suit. A chaque paire de mesures est associé un disque. Celui-ci est divisé en dix secteurs, chacun associé à un couple de base de cas et de règles. Le coefficient d'accord prenant ses valeurs entre -1 et 1, nous avons recentré ses valeurs : le rayon des portions de disques vaut $\frac{\tau_1+1}{2}$.

De plus, afin de mieux faire ressortir les structures de blocs, nous avons utilisé la méthode AMADO [Chauchat and Risson, 1998] sur la matrice moyenne du coefficient recentré.

3 Mise en regard des résultats théoriques et expérimentaux

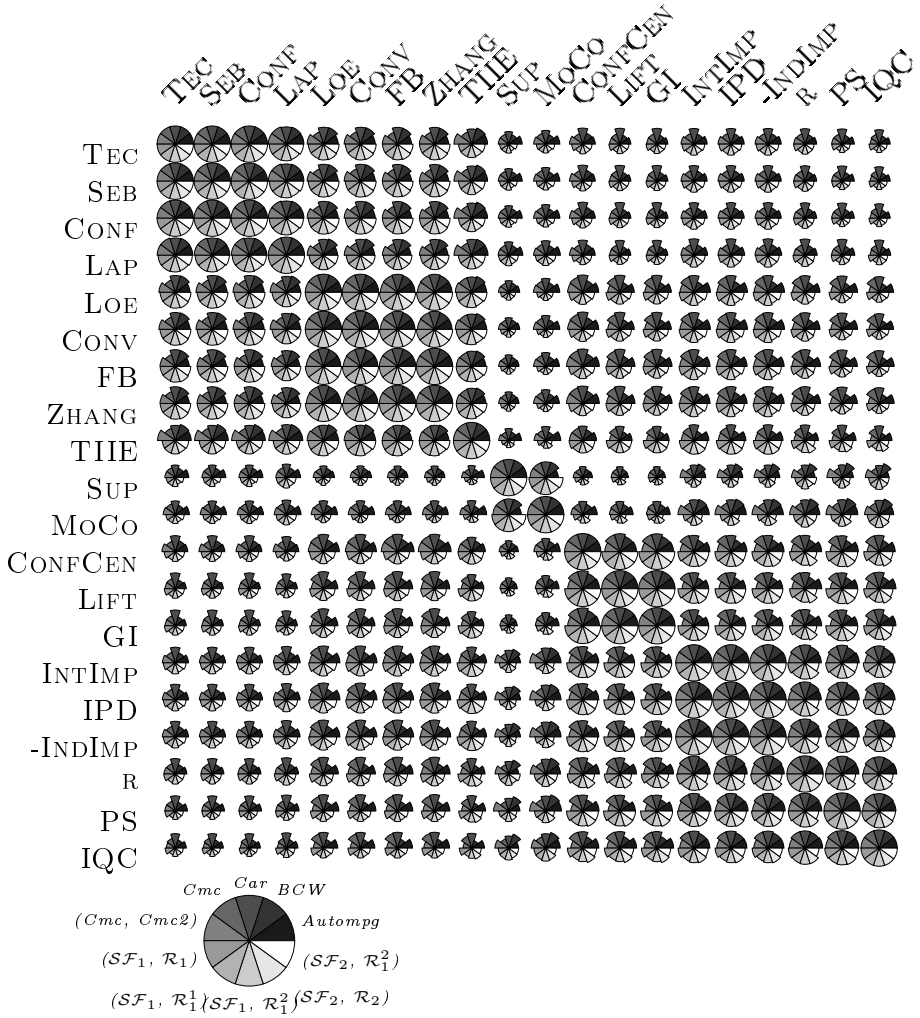
L'utilisation de la méthode AMADO permet de dégager de ces études expérimentales une classification des vingt mesures de qualité. De manière très distincte, trois classes principales se dégagent :

- {TEC, SEB, CONF, LAP, LOE, CONV, FB, ZHANG, TIIE },
- {SUP, MOCO },
- {CONFCEN, LIFT, GI, INTIMP, IPD, -INDIMP, R, PS, IQC }.

Au sein de ces classes, des regroupements plus fins semblent se dégager. Par exemple, le groupe {LOE, CONV, FB, ZHANG, TIIE} semble avoir un comportement légèrement différent de celui de {TEC, SEB, CONF, LAP}. De manière similaire, le comportement en situation de {CONFCEN, LIFT, GI} se démarque sensiblement de celui de {INTIMP, IPD, -INDIMP, R, PS, IQC}.

La table IV.8 croise ces cinq classes avec celles obtenues formellement en section III. Les classes expérimentales y sont notées e_i, les formelles f_i. De cette confrontation, se dégage une impression de cohérence entre résultats expérimentaux et théoriques.

Quelques mesures ont des comportements qui varient. LAP par exemple, de par sa définition, prend des valeurs très proches de celles de CONF en pratique. Son évaluation théorique, qui la rapproche de SUP et MOCO est ainsi revue par l'approche en situation. On voit également que TIIE, formellement relativement proche d'INTIMP, s'en distingue bien en situation. INTIMP est d'ailleurs la seule mesure prenant une valeur fixe à la règle logique et à l'indépendance se démarquant du groupe {LOE, CONV, FB, ZHANG, TIIE}. Il est intéressant de voir ainsi la contribution non-négligeable apportée par l'indice d'inclusion sur les mesures entropiques.



Tab. IV.7 – Comparaison des préordres induits par vingt mesures

	e_1	e_2	e_3	e_5	e_5
f_1	{Lap}		{Sup, MoCo}		
f_2	{Tec, Seb, Conf}				
f_3				{ConfCen, Lift, GI}	{-IndImp, R, PS, IQC}
f_4		{Loe, Conv, FB}			
f_5		{Zhang, TIIE}			{IntImp, IPD}

Tab. IV.8 – Comparaison des classifications formelles et expérimentales

4 Développements futurs

L'outil Herbs que nous avons développé nous a permis de mener des campagnes expérimentales sur un nombre conséquent de bases de règles, et nous avons présenté le résultat de la comparaison des préordres obtenu sur dix bases, pour vingt mesures.

Nous n'avons retenu qu'un seul coefficient d'accord, au vu des propriétés que nous souhaitions que celui-ci possède. Il est très aisé d'ajouter d'autres coefficients, tel le coefficient de corrélation par exemple.

Les besoins de visualisation et de représentation synthétique de l'information nous ont incité à implémenter des vues variées tant dans leurs natures (graphiques ou numériques) que dans leurs niveaux de précisions (ces niveaux de détail allant de l'analyse pour chaque règle à une synthèse sur l'ensemble global de règles).

L'intégration à Herbs de l'approche par agrégation est en cours, et permet d'obtenir une synthèse non plus sur l'ensemble des règles, mais cette fois sur un ensemble de mesures. Les méthodes de présentations du graphe de surclassement et le choix des opérateurs d'agrégations sont des sujets à approfondir. L'intégration d'autres méthodes d'aide à la sélection de mesures ou de règles est également envisagée.

Actuellement, Herbs intègre vingt-quatre des mesures que nous avons présenté au chapitre II. Compléter cette liste se fait très simplement pour les mesures descriptives. Il est également envisagé d'augmenter cette liste en incluant les versions généralisées des mesures. Toutefois, cette étape nécessite des modifications au niveau de la base de données, et reste à entreprentre.

Conclusion

La validation de connaissances extraites de données, en particulier dans le contexte des règles d'association, peut fortement reposer sur l'utilisation de mesures de qualité. Ces dernières ont pour objectif de mettre en avant ce qui est intéressant du point de vue d'un utilisateur expert, tant la grande quantité de règles pouvant être extraites rend l'évaluation exhaustive de celles-ci impossible en pratique.

Les mesures de qualité disponibles sont nombreuses et leurs caractéristiques tant formelles qu'en situation attestent de leur diversité. Ainsi, le choix d'utilisation d'une ou plusieurs mesures particulières lors de l'étape finale de validation a des conséquences sur le résultat du processus d'extraction de connaissances à partir de données (ECD). Se pose alors le problème du choix de mesures à prendre en compte.

Nos travaux répondent à cette problématique selon deux approches formelles et une étude expérimentale.

Nous avons ainsi étudié un nombre important de mesures de qualité de règles d'association. Parmi celles-ci, nous en avons retenu vingt, que nous avons caractérisées de manière transversale selon neuf propriétés pertinentes et adaptées aux problématiques abordées en ECD. A partir de ces propriétés, formulées en des termes les plus subjectifs possible, en prenant en compte les préférences de l'utilisateur, il nous est alors possible de mettre en œuvre un processus d'aide multicritère à la décision (AMD). Nous avons présenté les résultats obtenus avec la méthode PROMETHEE. Celle-ci propose un rangement des mesures. On peut alors y trouver celles répondant le mieux aux besoins exprimés. Une étude de la stabilité de ce rangement est également réalisée. La mise en œuvre d'autres méthodes d'AMD sur le même problème conforte nos résultats.

Cette étude transversale nous amène également à proposer une classification fondée sur cette analyse formelle de l'espace des mesures. Nous la mettons en regard d'une classification issue d'une comparaison expérimentale de leurs comportements.

Nous ne prétendons pas que cette étude soit exhaustive, mais elle constitue néanmoins un premier cadre de travail. Nous proposons en effet de systématiser l'étude de mesures de qualité selon des propriétés adaptées au contexte. La liste des mesures et des propriétés peut bien évidemment être étendue, sans remettre en cause la démarche. Elle peut également être adaptée à d'autres contextes, comme celui des arbres de décision.

Ainsi, la résistance des règles et des mesures au bruit contenu dans les données est un exemple de propriété qu'il nous semble pertinent d'explorer, au-delà de la modélisation simple que nous avons proposée. Notre modélisation a l'avantage d'être directe, et n'apporte

donc pas de surcoût de calcul. Toutefois, suite à cette première analyse, l'utilisateur pourrait souhaiter détailler son étude sur un nombre restreint d'éléments retenus.

Par ailleurs, nous avons développé une plate-forme, HERBS, permettant de mener des campagnes expérimentales en situation. Celle-ci nous donne les moyens d'analyser le comportement de mesures de qualité selon deux modes : l'un prospectif, en appliquant des traitements pas à pas, l'autre à grande échelle, par enchaînement ou sérialisation de ces derniers. La plate-forme peut également produire des résultats en situation sur un autre modèle de connaissance, les arbres de décision.

Qui plus est, lors de la construction du cadre formel, nous avons révisé certaines propriétés proposées dans la littérature et souligné l'importance que peut avoir le choix d'une modélisation particulière sur leur évaluation.

Cette construction a également guidé notre réflexion quant à la prise en compte d'une valeur de référence par les mesures. Ceci a débouché sur la proposition de familles de mesures généralisées. L'introduction d'un paramètre à définir par l'utilisateur, selon la logique de fouille qu'il adopte (à savoir *prédictive*, *de ciblage*, etc.), constitue un cadre unifiant sur les mesures de qualité.

La mise en œuvre de nos travaux, au delà des deux scenarii présentés et de travaux connexes aux nôtres, est en cours. Celle-ci s'effectue en collaboration avec un expert métier, et des premiers échanges ont permis de progresser dans les phases initiales de la fouille de données.

La seconde approche formelle que nous proposons pour résoudre le problème de choix apporte des éléments de réponses différents. Plutôt que de choisir un sous-ensemble restreint de mesures, on y réalise une synthèse des avis que les mesures expriment sur chaque règle, à savoir si elles la jugent meilleure que les autres ou non. Cette voie n'est pas purement ordinale, puisque l'intensité de préférence d'une règle sur une autre est prise en compte. Ainsi, l'on peut tenir compte des spécificités de chaque mesure (un petit écart d'évaluation pour l'une d'entre elles pouvant être tout aussi significatif qu'un grand pour une autre).

Nous avons montré qu'il était impossible de construire une relation agrégée vérifiant toutes les propriétés que nous souhaitions conserver (tout particulièrement la transitivité, et la prise en compte de l'ensemble des avis), pour les opérateurs étudiés. En effectuant des compromis, il est cependant possible d'obtenir des relations entre règles dont le comportement reste cohérent pour les situations peu conflictuelles.

Plusieurs pistes restent à explorer afin d'améliorer les résultats obtenus par cette approche. D'une part, l'étude d'autres opérateurs pourrait conduire à des agrégations respectant mieux nos souhaits. D'autre part, il est probablement possible de relâcher certaines contraintes en ayant une approche plus locale du problème.

Enfin, le choix de certains paramètres internes à la méthode utilisée méritent d'être étudié plus amplement. Les choix que nous avons effectués nous semblent de bon sens, et ont l'avantage d'être calculés directement en fonction des valeurs que prennent les mesures. En contre-partie, l'utilisateur maîtrise moins ces paramètres, ou pourrait souhaiter les positionner autrement.

Bibliographie

[Agrawal et al., 1993] Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In Buneman, P. and Jajodia, S., editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C. 11, 12, 47

[Agrawal and Srikant, 1994] Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In Bocca, J. B., Jarke, M., and Zaniolo, C., editors, *Proceedings of the 20th Very Large Data Bases Conference*, pages 487–499. Morgan Kaufmann. 2, 17, 18

[Arrow, 1951] Arrow, K. J. (1951). *Social Choice and Individual Values*. Cowles Foundations and Wiley. 72

[Azé, 2003] Azé, J. (2003). *Extraction de Connaissances dans des Données Numériques et Textuelles*. PhD thesis, Université de Paris 11. 94

[Azé and Kodratoff, 2002] Azé, J. and Kodratoff, Y. (2002). Evaluation de la résistance au bruit de quelques mesures d'extraction de règles d'assocation. *Extraction des connaissances et apprentissage (Extraction et Gestion des Connaissances 2002)*, 1(4) :143–154. 35, 47

[Badjio and Poulet, 2006] Badjio, E. P. F. and Poulet, F. (2006). Prétraitement de grands ensembles de données pour la fouille visuelle. In *Extraction et Gestion des Connaissances*, pages 59–64. 66

[Barthélemy et al., 2006] Barthélemy, J. P., Legrain, A., Lenca, P., and Vaillant, B. (2006). Aggregation of valued relations applied to association rule interestingness measures. In Torra, V., Narukawa, Y., Valls, A., and Domingo-Ferrer, J., editors, *Modeling Decisions for Artificial Intelligence*, volume 3885 of *Lecture Notes in Computer Science*, pages 203–214, Tarrogona, Spain. Springer-Verlag. 49, 65, 101

[Bellman and Zadeh, 1970] Bellman, R. and Zadeh, L. (1970). Decision-making in a fuzzy environment. *Management of Sciences 17*, pages 141–164. 73

[Berti-Équille, 2004] Berti-Équille, L. (2004). La qualité des données comme condition à la qualité des connaissances : un état de l'art. *Revue des Nouvelles Technologies de l'Information (Mesures de Qualité pour la Fouille de Données)*, (RNTI-E-1). 136

[Bisdorff, 2004] Bisdorff, R. (2004). Kernels from di-graphs : applications to selecting, ranking and clustering problems. invited plenary lecture. In Society, S.-B. B. O. R., editor, *18th Conference on Quantitative Methods for Decision Making*, Brussels. Royal Military Academy. 64

[Blanchard, 2005] Blanchard, J. (2005). *Un système de visualisation pour l'extraction, l'évaluation, et l'exploration interactives des règles d'association*. PhD thesis, Ecole Polytechnique de l'Université de Nantes. 19

[Blanchard et al., 2003] Blanchard, J., Guillet, F., and Briand, H. (2003). Exploratory visualization for association rule rummaging. In *Proceedings of the KDD'2003 Workshop on Multimedia Data Mining MDM'03*, pages 107–114. 66

[Blanchard et al., 2005a] Blanchard, J., Guillet, F., Briand, H., and Gras, R. (2005a). IPEE : Indice Probabiliste d'Ecart à l'Equilibre pour l' 'evaluation de la qualité des règles. In *Atelier Qualité des Données et des Connaissances*, pages 26–34. 25, 44, 47, 54, 82, 85

[Blanchard et al., 2005b] Blanchard, J., Guillet, F., Briand, H., and Gras, R. (2005b). Une version discriminante de l'indice probabiliste d'écart à l'équilibre pour mesurer la qualité des règles. In *Troisièmes rencontres internationales de l'Analyse Statistique Implicative (ASI 05)*. 45, 47

[Blanchard et al., 2004] Blanchard, J., Kuntz, P., Guillet, F., and Gras, R. (2004). Mesure de la qualité des règles d'association par l'intensité d'implication entropique. *Revue des Nouvelles Technologies de l'Information (Mesures de Qualité pour la Fouille de Données)*, (RNTI-E-1) :33–44. 55

[Borda, 1781] Borda, J. (1781). *Mémoire sur les Élections au Scrutin*. Comptes rendus de l'Académie des sciences. 73

[Borgelt and Kruse, 2002] Borgelt, C. and Kruse, R. (2002). Induction of association rules : APRIORI implementation. In *Proceedings of the 15th Conference on Computational Statistics*, Heidelberg, Germany. Physika Verlag. 17, 27

[Brans and Mareschal, 1994] Brans, J. and Mareschal, B. (1994). The PROMETHEE-GAIA decision support system for multicriteria investigations. *Investigation Operativa*, 4(2) :102–117. 66

[Brans and Mareschal, 2002] Brans, J. and Mareschal, B. (2002). PROMETHEE-GAIA – *Une Méthode d'Aide à la Décision en Présence de Critères Multiples*. Ellipses. 66

[Brans et al., 1984] Brans, J.-P., Mareschal, B., and Vincke, P. (1984). PROMETHEE *: A New Family of Outranking Methods in MCDM*. IFORS 84. 64, 66, 104

[Breiman et al., 1984] Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth, Belmont. 23, 31

[Brin et al., 1997a] Brin, S., Motwani, R., and Silverstein, C. (1997a). Beyond market baskets : generalizing association rules to correlations. In *ACM SIGMOD/PODS '97 Joint Conference*, pages 265–276. 29, 32, 47

[Brin et al., 1997b] Brin, S., Motwani, R., Ullman, J. D., and Tsur, S. (1997b). Dynamic itemset counting and implication rules for market basket data. In Peckham, J., editor, *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, Tucson, Arizona, USA*, pages 255–264. ACM Press. 18, 47

[Carvalho et al., 2005] Carvalho, D., Freitas, A. A., and Ebecken, N. (2005). Evaluating the correlation between objective rule interestingness measures and real human interest. In Jorge, A., L., T., Brazdil, P., Camacho, R., and Gama, J., editors, *Knowledge Discovery in Databases : Proceedings of PKDD 2005, LNAI 3731*, pages 453–461. Springer Verlag. 2

[Chauchat and Risson, 1998] Chauchat, J.-H. and Risson, A. (1998). *Visualization of Categorical Data*, chapter 3, pages 37–45. Blasius J. & Greenacre M. ed. New York : Academic Press. 139

[Choquet, 1953] Choquet, G. (1953). *Theory of Capacities*, pages 131–295. Annales de l'Institut Fourier (Grenoble). 105

[Church and Hanks, 1990] Church, K. W. and Hanks, P. (1990). Word association norms, mutual information an lexicography. *Computational Linguistics*, 16(1) :22–29. 31, 47

[Clark and Niblett, 1989] Clark, P. and Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning*, 3(4) :261–283. 30

[Cohen, 1960] Cohen, J. (1960). A coefficient of agreement for nominal scale. *Educational and Psychological Measurement*, 20 :37–46. 32, 47

[Condorcet, 1785] Condorcet, M. (1785). *Essai sur l'Application de l'Analyse à la Probabilité des Décisions Rendues à la Pluralité des Voix*. Imprimerie Royale, Paris. 73

[Cornelis et al., 2006] Cornelis, C., Yan, P., Zhang, X., and Chen, G. q. (2006). Mining positive and negative association rules from large databases. In *The IEEE International Conference on Cybernetics and Intelligent Systems*, pages 613–618, Bangkok, Thailand. 19

[Do and Poulet, 2003] Do, T.-N. and Poulet, F. (2003). Interactive visualization tools for visual data-mining. In Bisdorff, R., editor, *Human Centered Processes (HCP'2003)*, pages 299–303, Luxembourg. 19

[Dong et al., 2006] Dong, X., Sun, F., Han, X., and Hou, R. (2006). Study of positive and negative association rules based on multi-confidence and chi-squared test. In Li, X., Zaiane, O., and Li, X., editors, *Advanced Data Mining and Applications*, volume 4093 of *Lecture Notes in Artificial Intelligence*, pages 100–109, Xi'an, China. Springer-Verlag. 19

[Fayyad et al., 1996] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors (1996). *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press. 1, 5

[Fodor and Roubens, 1994] Fodor, J. and Roubens, M. (1994). *Fuzzy preference modelling and multicriteria decision support*. Kluwer Academic Publishers. 67, 75, 79

[Fodor and Roubens, 1995] Fodor, J. and Roubens, M. (1995). Structure of transitive valued binary relations. *Mathematical social sciences*, 30 :71–94. 77

[Francisci et al., 2003] Francisci, D., Brisson, L., and Collard, M. (2003). Extraction de règles selon des critères multiples : l'art du compromis. Technical Report ISRN I3S/RR-2003-11-FR, Université de Nice. 119

[Freitas, 1999] Freitas, A. (1999). On rule interestingness measures. *Knowledge-Based Systems journal*, pages 309–315. 21, 82

[Fukuda et al., 1996] Fukuda, T., Morimoto, Y., Morishita, S., and Tokuyama, T. (1996). Data mining using two-dimensional optimized association rules : scheme, algorithms, and visualization. In *ACM SIGMOD International Conference on Management of Data*, pages 13–23. 32, 47

[Ganascia, 1987] Ganascia, J.-G. (1987). CHARADE : A rule system learning system. In *10th International Joint Conference on Artificial Intelligence*, Milan. 33, 47

[Gardarin et al., 1998] Gardarin, G., Pucheral, P., and Wu, F. (1998). Bitmap based algorithms for mining association rules. Technical report, Université de Versailles Saint-Quentin. 19

[Giakoumakis and Monjardet, 1987] Giakoumakis, V. and Monjardet, B. (1987). Coefficients d'accords entre deux préordres totaux. *Statistique et analyse de données.* 133

[Gini, 1912] Gini, C. (1912). Variability and mutability. In *Contribution to the study of statistical distributions and relations.* Studi Economico-Giuridici della R. Universita de Cagliari. 23

[Good, 1965] Good, I. J. (1965). The estimation of probabilities : An essay on modern bayesian methods. The MIT Press, Cambridge, MA. 30, 47

[Goodman and Smyth, 1988] Goodman, R. M. and Smyth, P. (1988). Information-theoretic rule induction. In *Proceedings of the 1988 European Conference on Artificial Intelligence.* Pitman. 23

[Grabisch and Perny, 2003] Grabisch, M. and Perny, P. (2003). Agrégation multicritère. In Bouchon-Meunier, B. and Marsala, C. e., editors, *In Logique floue, principes, aide à la décision*, pages 81–120. Hermès. 73, 79

[Gras, 1979] Gras, R. (1979). *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques.* PhD thesis, Université de Rennes I. 39, 47, 54, 90

[Gras et al., 1996] Gras, R., Ag. Almouloud, S., Bailleuil, M., Larher, A., Polo, M., Ratsimba-Rajohn, H., and Totohasina, A. (1996). *L'implication Statistique, Nouvelle Méthode Exploratoire de Données. Application à la Didactique, Travaux et Thèses.* La Pensée Sauvage. 39, 47

[Gras et al., 2004] Gras, R., Couturier, R., Blanchard, J., Briand, H., Kuntz, P., and Peter, P. (2004). Quelques critères pour une mesure de qualité de règles d'association - un exemple : l'intensité d'implication. *Revue des Nouvelles Technologies de l'Information (Mesures de Qualité pour la Fouille de Données)*, (RNTI-E-1) :3–31. 82, 90

[Gras et al., 2001] Gras, R., Kuntz, P., Couturier, R., and Guillet, F. (2001). Une version entropique de l'intensité d'implication pour les corpus volumineux. *Extraction des connaissances et apprentissage (Extraction et Gestion des Connaissances 2001)*, 1(1-2) :69–80. 41, 47, 55, 56, 82, 86

[Greco et al., 2002] Greco, S., Matarazzo, B., and Slowinski, R. (2002). Data mining tasks and methods : Classification : multicriteria classification. pages 318–328. 67

[Guillaume, 2000] Guillaume, S. (2000). *Traitement des données volumineuses : mesures et algorithmes d'extraction de règles d'association et règles ordinales.* PhD thesis, Université de Nantes. 19

[Guillet, 2004] Guillet, F. (2004). Mesure de la qualité des connaissances en ECD. Tutoriel de la 4e Conf. Extraction et Gestion des Connaissances. 60 pages. 21

[Hilderman and Hamilton, 2003] Hilderman, R. and Hamilton, H. (2003). Measuring the interestingness of discovered knowledge : A principled approach. *Intelligent Data Analysis*, 7(4) :347–382. 21, 81

[Hilderman and Hamilton, 2000] Hilderman, R. J. and Hamilton, H. J. (2000). Applying objective interestingness measures in data mining systems. In *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'00)*, pages 432–439. Springer-Verlag. 19, 21

[Hilderman and Hamilton, 2001a] Hilderman, R. J. and Hamilton, H. J. (2001a). Evaluation of interestingness measures for ranking discovered knowledge. *Lecture Notes in Computer Science*, 2035 :247–259. 21

[Hilderman and Hamilton, 2001b] Hilderman, R. J. and Hamilton, H. J. (2001b). *Knowledge Discovery and Measures of Interest*, volume 638 of *The International Series in Engineering and Computer Science*. Kluwer. 2, 81

[Hofmann and Wilhelm, 2001] Hofmann, H. and Wilhelm, A. (2001). Visual comparison of association rules. *Computational Statistics*, 16(3) :399–415. 19

[Holt and Chung, 2002] Holt, J. D. and Chung, S. M. (2002). Mining association rules using inverted hashing and pruning. *Information Processing Letter*, 83(4) :211–220. 18

[Huynh et al., 2006] Huynh, X.-H., Guillet, F., and Briand, H. (2006). ARQAT : plateforme exploratoire pour la qualité des règles d'association. *Revue des Nouvelles Technologies de l'Information (Extraction et gestion des connaissances : Etat et perspectives)*, (RNTI-E-5). 121

[Jaccard, 1908] Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin Society Vaud Science National*, (44) :223–270. 34, 47

[Jeffreys, 1935] Jeffreys, H. (1935). Some tests of significance treated by the theory of probability. In *Proceedings of the Cambridge Philosophical Society*, pages 203–222. 30, 47

[Kelly, 1991] Kelly, J. S. (1991). Social choice bibliography. *Social Choice and Welfare*, 8 : 97-169. 72

[Kodratoff et al., 2001] Kodratoff, Y., Napoli, A., and Zighed, D. (2001). Bulletin de l'association française d'intelligence artificielle, extraction de connaissances dans des bases de données. 1, 5

[Kuntz et al., 2001] Kuntz, P., Lehn, R., Guillet, F., and Briand, H. (2001). Vers un processus d'extraction de règles d'association centré sur l'utilisateur. *InCognito*, (20) :13–26. 19

[Kuntz et al., 2004] Kuntz, P., Pinaud, B., and Lehn, R. (2004). Elements for the description of fitness landscapes associated with local operators for layered drawings of directed graphs. In *Metaheuristics : computer decision-making*, pages 405–420. Kluwer Academic Publishers, Norwell, MA, USA. 119

[Lallich, 2002] Lallich, S. (2002). Mesure et validation en extraction des connaissances à partir des données. Habilitation à Diriger des Recherches – Université Lyon 2. 82, 90

[Lallich et al., 2005a] Lallich, S., Lenca, P., and Vaillant, B. (2005a). Variations autour de l'intensité d'implication. In *The Third International A.S.I. Conference Implicative Statistic Analysis*, pages 237–246. 44

[Lallich et al., 2004] Lallich, S., Prudhomme, E., and Teytaud, O. (2004). Contrôle du risque multiple en sélection de règles d'association significatives. *RNTI-E-2 (Extraction et Gestion des Connaissances 2004)*, 2 :305–316. 138

[Lallich and Teytaud, 2004] Lallich, S. and Teytaud, O. (2004). Évaluation et validation de l'intérêt des règles d'association. *Revue des Nouvelles Technologies de l'Information*. 82, 90, 95

[Lallich et al., 2005b] Lallich, S., Vaillant, B., and Lenca, P. (2005b). Parametrised measures for the evaluation of association rule interestingness. In *The XIth International Symposium on Applied Stochastic Models and Data Analysis*, pages 220–229, Brest, France. 42, 47, 49, 52, 53, 55

[Lallich et al., 2007] Lallich, S., Vaillant, B., and Lenca, P. (scheduled for June 2007). A probabilistic framework towards the parameterization of association rule interestingness measures. *Methodology and Computing in Applied Probability*. 49

[Leclerc, 1984] Leclerc, B. (1984). Efficient and binary consensus functions on transitively valued relations. *Mathematical Social Sciences*, 8. 80

[Legrain, 2004] Legrain, A. (2004). Agrégation de mesures de qualité de règles d'association. Master's thesis, Ecole Nationale Supérieure des Télécommunications de Bretagne. 65, 75, 78, 79, 105, 119

[Lehn, 2000] Lehn, R. (2000). *Un système interactif de visualisation et de fouille de règles pour l'extraction de connaissances dans les bases de données*. PhD thesis, Université de Nantes. 19, 66, 119

[Lehn et al., 1998] Lehn, R., Guillet, F., and Briand, H. (1998). Eliminating redundancy in a rule system. In *Eurepean meeting on Cybernetics and System Research*, volume 2, pages 793–798. 19

[Lehn et al., 1999] Lehn, R., Guillet, F., Kuntz, P., Briand, H., and Philippé, J. (1999). Felix : An interactive rule mining interface in a kdd process. In Lenca, P., editor, *Proceedings of the Human Centered Processes Conference*, pages 169–174, Brest, France. 19

[Lenca et al., 2003a] Lenca, P., Meyer, P., Picouet, P., Vaillant, B., and Lallich, S. (2003a). Critères d'évaluation des mesures de qualité en ECD. *Revue des Nouvelles Technologies de l'Information (Entreposage et Fouille de données)*, (1) :123–134. 2, 82, 84, 95

[Lenca et al., 2003b] Lenca, P., Meyer, P., Vaillant, B., and Picouet, P. (2003b). Aide multi-critère à la décision pour évaluer les indices de qualité des connaissances – modélisation des préférences de l'utilisateur. *RSTI-RIA (Extraction et Gestion des Connaissances 2003)*, 1(17) :271–282. 21, 63, 64, 82, 90

[Lenca et al., 2006] Lenca, P., Vaillant, B., and Lallich, S. (2006). On the robustness of association rules. In *IEEE International Conference on Cybernetics and Intelligent Systems*, pages 596–601, Bangkok, Thailand. 90, 91, 92

[Lerman and Azé, 2003] Lerman, I. and Azé, J. (2003). Une mesure probabiliste contextuelle discriminante de qualité des règles d'association. *RSTI-RIA (Extraction et Gestion des Connaissances 2003)*, 1(17) :247–262. 44, 47, 55, 56

[Lerman et al., 1981] Lerman, I., Gras, R., and Rostam, H. (1981). Elaboration d'un indice d'implication pour les données binaires, i et ii. *Mathématiques et Sciences Humaines*, (74, 75) :5–35, 5–47. 37, 38, 39, 53, 54, 95

[Lin and Tseng, 2006] Lin, W.-Y. and Tseng, M.-C. (2006). Automated support specification for efficient mining of interesting association rules. *Journal of Information Science*, 32(3) :238–250. 27

[Liu et al., 1997] Liu, B., Hsu, W., and Chen, S. (1997). Using general impressions to analyze discovered classification rules. In *ACM International Conference on Knowledge Discovery and Data Mining*, pages 31–36. ACM Press. 22

[Liu et al., 2000] Liu, B., Hsu, W., Chen, S., and Ma, Y. (2000). Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems*, 15(5) :47–55. 21

[Loevinger, 1947] Loevinger, J. (1947). A systemic approach to the construction and evaluation of tests of ability. *Psychological monographs*, 61(4). 28, 47

[Lukasiewicz, 1970] Lukasiewicz, J. (1970). *selected works*. Reidel, Amsterdam. 76

[Mannila et al., 1994] Mannila, H., Toivonen, H., and Verkamo, A. I. (1994). Efficient algorithms for discovering association rules. In Fayyad, U. M. and Uthurusamy, R., editors, *AAAI Workshop on Knowledge Discovery in Databases (KDD-94)*, pages 181–192, Seattle, Washington. AAAI Press. 18

[Marichal et al., 2005] Marichal, J.-L., Meyer, P., and Roubens, M. (2005). Sorting multi-attribute alternatives : The TOMASO method. *Computers & Operations Research*, (32) :861–877. 65

[McGarry, 2005] McGarry, K. (2005). A survey of interestingness measures for knowledge discovery. *Knowledge Engineering Review*, 20(1) :39–61. 21

[Newman et al., 1998] Newman, D., Hettich, S., Blake, C., and Merz, C. (1998). UCI repository of machine learning databases. 123, 135

[Padmanabhan and Tuzhilin, 1998] Padmanabhan, B. and Tuzhilin, A. (1998). A belief-driven method for discovering unexpected patterns. In *ACM International Conference on Knowledge Discovery and Data Mining*, pages 94–100. ACM Press. 22

[Pagès and Groetaers, 2003] Pagès, N. and Groetaers, D. N.-F. (2003). Visualisation of usability data. In Bisdorff, R., editor, *Human Centered Processes*, pages 293–298, Luxembourg. 19

[Park et al., 1997] Park, J. S., Chen, M.-S., and Yu, P. S. (1997). Using a hash-based method with transaction trimming for mining association rules. *IEEE Transactions on Knowledge and Data Engineering*, 9(5) :813–825. 18

[Park et al., 1995] Park, J. S., M.-S., C., and Yu, P. S. (1995). An effective hash-based algorithm for mining association rules. In *SIGMOD '95 : Proceedings of the 1995 ACM SIGMOD international conference on Management of data*, pages 175–186, New York, NY, USA. ACM Press. 18

[Pearson, 1896] Pearson, K. (1896). Mathematical contributions to the theory of evolution. iii. regression, heredity and panmixia. *Philosophical Transactions of the Royal Society*, A. 28, 47

[Peneva and Popchev, 2003] Peneva, V. and Popchev, I. (2003). Properties of the aggregation operators related with fuzzy relations. *Fuzzy sets and systems*, pages 615–633. 80

[Piatetsky-Shapiro, 1991] Piatetsky-Shapiro, G. (1991). Discovery, analysis and presentation of strong rules. In Piatetsky-Shapiro, G. and Frawley, W., editors, *Knowledge Discovery in Databases*, pages 229–248. AAAI/MIT Press. 35, 47, 82, 83, 85

[Plasse et al., 2006] Plasse, M., Niang, N., Saporta, G., and Leblond, L. (2006). Une comparaison de certains indices de pertinence des règles d'association. In *Extraction et Gestion des Connaissances*, pages 561–568. 101

[Quinlan, 1992] Quinlan, J. R. (1992). *C4.5 : Programs for Machine Learning*. Morgan Kaufmann. 23

[Roy, 1968] Roy, B. (1968). Classement et choix en présence de points de vue multiples (la méthode electre). *Revue d'Informatique et de Recherche Opérationnelle*, (8) :57–75. 64

[Roy, 1985] Roy, B. (1985). *Méthodologie multicritère d'aide à la décision*. Collection Gestion. Economica. 64

[Roy and Bouyssou, 1993] Roy, B. and Bouyssou, D. (1993). *Aide Multicritère à la Décision : Méthodes et Cas*. Economica. 64, 72, 73

[Savasere et al., 1995] Savasere, A., Omiecinski, E., and Navathe, S. B. (1995). An efficient algorithm for mining association rules in large databases. In *Proceedings of the 21th Very Large Data Bases Conference*, pages 432–444. 19

[Sebag and Schoenauer, 1988] Sebag, M. and Schoenauer, M. (1988). Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases. In Boose, J., Gaines, B., and Linster, M., editors, *Proc. of the European Knowledge Acquisition Workshop (EKAW'88)*, pages 28–1 – 28–20. Gesellschaft für Mathematik und Datenverarbeitung mbH. 34, 47

[Sese and Morishita, 2002] Sese, J. and Morishita, S. (2002). Answering the most correlated n association rules efficiently. In *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 410–422. Springer-Verlag. 27

[Silberschatz and Tuzhilin, 1995] Silberschatz, A. and Tuzhilin, A. (1995). On subjective measures of interestingness in knowledge discovery. In *Knowledge Discovery and Data Mining*, pages 275–281. 21, 22

[Sugeno, 1974] Sugeno, M. (1974). *The Theory of Fuzzy Integrals and Its Applications*. PhD thesis, Tokyo Institue of Technology. 105

[Szathmary et al., 2006] Szathmary, L., Maumus, S., Petronin, P., Toussaint, Y., and Napoli, A. (2006). Vers l'extraction de motifs rares. In *Extraction et Gestion des Connaissances*, pages 499–510. 11

[Tan et al., 2002] Tan, P.-N., Kumar, V., and Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. In *Proceedings of the Eighth ACM SIGKDD International Conference on KDD*, pages 32–41. 2, 21, 23, 81

[Tversky and Russo, 1969] Tversky, A. and Russo, J. (1969). Substituability and similarity in binary choices. *Journal of Mathematical Psychology*, 6 :1–12. 77

[Vaillant, 2002] Vaillant, B. (2002). Evaluation de connaissances : le problème du choix d'une mesure de qualité en extraction de connaissances à partir des données. Master's thesis, Ecole Nationale Supérieure des Télécommunications de Bretagne. 121

[Vaillant et al., 2006a] Vaillant, B., Lallich, S., and Lenca, P. (2006a). Modeling of the counter-examples and association rules interestingness measures behavior. In *The 2006 International Conference on Data Mining*, pages 132–137, Las Vegas, Nevada, USA. 23, 87, 89

[Vaillant et al., 2004] Vaillant, B., Lenca, P., and Lallich, S. (2004). A clustering of interestingness measures. In *Discovery Science*, pages 290–297. 96, 136, 137

[Vaillant et al., 2006b] Vaillant, B., Meyer, P., Prudhomme, E., Lallich, S., Lenca, P., and Bigaret, S. (2006b). Mesurer l'intérêt des règles d'association. *Revue des Nouvelles Technologies de l'Information (Extraction et gestion des connaissances : Etat et perspectives)*, (RNTI-E-5) :421–426. 96, 121

[Vaillant et al., 2002] Vaillant, B., Picouet, P., and Lenca, P. (2002). Une plateforme extensible pour l'évaluation de mesures de qualité de règles. Rapport de recherche pour le groupe de travail GAFOQUALITÉ de l'action spécifique STIC fouille de bases de données, Département IASC, ENST Bretagne. 121

[Vaillant et al., 2003] Vaillant, B., Picouet, P., and Lenca, P. (2003). An extensible platform for rule quality measure benchmarking. In Bisdorff, R., editor, *Human Centered Processes (HCP'2003)*, pages 187–191, Luxembourg. 121

[Vincke, 1989] Vincke, P. (1989). *L'aide multicritère à la décision.* Editions de l'Université de Bruxelles. 64, 67, 72

[Wu et al., 2004] Wu, X., Zhang, C., and Zhang, S. (2004). Efficient mining of both positive and negative association rules. *ACM Transactions on Information Systems*, 22(3) :381–405. 19

[Yager, 1988] Yager, R. R. (1988). On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Transactions on Systems, Man and Cybernetics*, 18(1) :183–190. 105

[Yao and Zhong, 1999] Yao, Y. Y. and Zhong, N. (1999). An analysis of quantitative measures associated with rules. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 479–488. 21

[Zadeh, 1965] Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8 :338–353. 76

[Zaki, 2000] Zaki, M. (2000). Generating non-redundant association rules. In *Knowledge Discovery and Data Mining*, pages 34–43. 19

[Zaki et al., 1997] Zaki, M. J., Parthasarathy, S., Ogihara, M., and Li, W. (1997). Parallel algorithms for discovery of association rules. *Data Mining and Knowledge Discovery*, 1(4) :343–373. 19

[Zaman Ashrafi et al., 2004] Zaman Ashrafi, M., Taniar, D., and Smith, K. (2004). A new approach of eliminating redundant association rules. In Galindo, F. and Takizawa, M. Traunmüller, R., editors, *Database and Expert Systems Applications*, volume 3180 of *Lecture Notes in Computer Science*, pages 465–474, Zaragoza, Spain. Springer. 19

[Zhang, 2000] Zhang, T. (2000). Association rules. In Terano, T., Liu, H., and Chen, A., editors, *4th Pacific-Asia Conference Knowledge Discovery and Data Mining, Current Issues and New Applications*, volume 1805 of *Lecture Notes in Computer Science*, Kyoto, Japan. Springer. 37, 47

[Zighed and Rakotomalala, 2000] Zighed, D. A. and Rakotomalala, R. (2000). *Graphes d'induction : apprentissage et data mining.* Hermes Science Publications, Paris. 23

[Zimmermann and Raedt, 2004] Zimmermann, A. and Raedt, L. D. (2004). Corclass : Correlated association rule mining for classification. In *Discovery Science*, pages 60–72. 27

Summary : Evaluating the interestingness of association rules through the use of interestingness measures : formal and experimental studies

Knowledge discovery in databases aims at extracting information contained in data warehouses. It is a complex process, in which several experts (those acquainted with data, analysts, processing specialists, etc.) must act together in order to reveal patterns, which will be evaluated according to several criteria : validity, novelty, understandability, exploitability, etc. Depending on the application field, these criteria may be related to differing concepts. In addition, constant improvements made in the methodological and technical aspects of data mining allow one to deal with ever-increasing databases. The number of extracted patterns follows the same increasing trend, without them all being valid, however. It is commonly assumed that the validation of the knowledge mined cannot be performed by a decision maker, usually in charge of this step in the process, without some automated help.

In order to carry out this final validation task, a typical approach relies on the use of functions which numerically quantify the pertinence of the patterns. Since such functions, called interestingness measures, imply an order on the patterns, they highlight some specific kind of information. Many measures have been proposed, each of them being related to a particular category of situations.

We here address the issue of evaluating the objective interestingness of the particular type of patterns that are association rules, through the use of such measures. Considering that the selection of "good" rules implies the use of appropriated measures, we propose a systematic study of the latter, based on formal properties expressed in the most straightforward terms. From this study, we obtain a clustering of many commonly-used measures which we confront with an experimental approach obtained by comparing the rankings induced by these measures on classical datasets. Analysing these properties enabled us to highlight some particularities of the measures. We deduce a generalised framework that includes a large majority of them.

We also apply two Multicriteria Decision Aiding methods in order to solve the issue of retaining pertinent rules. The first approach takes into account a modelling of the preferences expressed by an expert in the field being mined about the previously defined properties. From this modelling, we establish which measures are the most adapted to the specific context. The second approach addresses the problem of taking into account the potentially differing values that the measures take, and builds an aggregated view of the ordering of the rules by taking into account the differences in evaluations. These methods are applied to practical situations.

This work also led us to develop powerful dedicated software, Herbs. We present the processing it allows for rule selection purposes, as well as for the analysis of the behaviour of measures and visualisation aspects.

Without any claim to exhaustiveness in our study, the methodology we propose can be extended to new measures or properties, and is applicable to other data mining contexts.

Key-words : knowledge discovery in databases, association rule, interestingness measure, multicriteria decision aiding.

Résumé : Mesurer la qualité des règles d'association : études formelles et expérimentales

L'extraction de connaissances à partir de données vise à extraire des motifs contenus dans des entrepôts de données, dont la taille ne cesse de croître. C'est un processus complexe dans lequel plusieurs experts (métier, analyste, informaticien...) doivent coopérer afin de révéler des motifs qui seront évalués selon différents critères : validités, nouveauté, compréhensibilité, exploitabilité, etc. Or, ces dernier critères sont fortement liés au contexte applicatif et peuvent être formulés de différentes manières. De plus, les progrès techniques et méthodologiques permanents permettent de traiter des volumes de données de plus en plus importants. Par voie de conséquence, le nombre de motifs extraits est également de plus en plus grand, sans pour autant qu'ils soient tous valides – bien au contraire. Force est de constater que la validation de connaissances ne peut maintenant plus se faire sans assistance au décideur, ce dernier ayant bien souvent comme tâche la validation des motifs.

Afin de permettre de procéder à cette tâche finale de validation lors du processus, un moyen couramment retenu est l'utilisation de fonctions quantifiant numériquement la pertinence des connaissances. De telles fonctions, dites mesures de qualité, mettent en avant une typologie de connaissance donnée, en induisant un ordre sur celles-ci. De nombreuses mesures ont été proposées, chacune étant liée à des situations précises.

Nous adressons dans ce mémoire la problématique de l'évaluation objective de la qualité d'un type de motif particulier, les règles d'association, par de telles mesures. Considérant que la sélection des "bonnes" règles repose sur l'utilisation d'une mesure adaptée, nous proposons une étude systématique de ces dernières, basée sur une analyse de propriétés formelles, énoncées selon des termes les plus compréhensibles possibles. De cette étude, on produit une classification d'un nombre important de mesures classiques, que nous confrontons à une classification expérimentale, obtenue en comparant les rangements induits par les mesures pour plusieurs jeux de données. L'étude des propriétés classiques et la définition de nouvelles nous a également permis de mettre en avant certaines particularités des mesures. Nous en déduisons un cadre généralisant la plupart d'entre elles.

Nous appliquons également deux méthodes d'Aide Multicritère à la Décision afin de résoudre le problème de la sélection des règles pertinentes. La première approche se fonde sur la prise en compte d'un système de préférences exprimé par un expert de données sur les propriétés précédemment définies, en vue d'orienter le choix de mesure(s) adaptée(s) au contexte applicatif. La seconde approche adresse le problème de la prise en compte des valeurs potentiellement discordantes qu'expriment les mesures sur l'intérêt des règles, et construisons une vue agrégée de leur ordonnancement, en prenant en compte les écarts d'évaluation. Ces méthodes sont appliquées à des situations pratiques.

Ce travail nous a aussi conduit au développement d'un outil spécifique performant, HERBS. Nous présentons les traitements qu'il permet d'effectuer, tant en termes de sélection de règles, d'analyse du comportement des mesures et de visualisation.

Sans prétendre à l'exhaustivité de notre étude, la méthodologie proposée est cependant extensible, en termes de mesures et de propriétés, ainsi qu'appliquable à d'autres contextes de fouille de données.

mots clés : extraction de connaissances à partir de données, règle d'association, mesure de qualité, aide multicritère à la décision.