République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université A.MIRA-BEJAIA
Faculté des Sciences Exactes
Département d'Informatique
Laboratoire d'Informatique Médicale (LIMED)

THÈSE

Présentée par

Souhila GHANEM

Pour l'obtention du grade de

DOCTEUR EN SCIENCES

Filière : Informatique

Option : Réseaux et Systèmes Distribués

Thème

Des Contributions autour de l'Analyse Statistique Implicative

Soutenue le : 12/07/2022 Devant le Jury composé de :

Nom et Prénom	Grade		
Mme. Houda EL BOUHISSI	M-C-A	Univ. de Béjaïa	Présidente
Mr Raphaël COUTURIER	Professeur	IUT Belfort-Montbéliard	Rapporteur
Mr Sofiane AISSANI	M-C-A	Univ. de Béjaïa	Examinateur
Mme Djamila BOULAHROUZ	M-C-A	Univ. de Béjaïa	Examinatrice
Mr Zakaria EL BERRICHI	Professeur	Univ. de Sidi-Bel-Abbès	Examinateur
Mr Sofiane BOUKLI HACENE	Professeur	Univ. de Sidi-Bel-Abbès	Examinateur
Mr Pablo GREGORI HUERTA	A.Professeur	Univ. d'Espagne	Invité

Année Universitaire : 2021/2022

⁂ *Remerciements* ⁂

Merci mon Dieu, de m'avoir donné la force et le courage pour mener ce travail a terme.

Ma profonde et sincère gratitude va d'abord à Monsieur le Professeur Raphaël Couturier qui m'a honoré de sa confiance en acceptant de travailler en ma compagnie. Promoteur de ce mémoire, il m'a beaucoup aidé par la qualité de son encadrement particulier, par la rigueur de ses orientations scientifiques qui impose beaucoup d'estime et de respect. J'ai beaucoup bénéficié de ses connaissances étendues et compétences incroyables. Outre son appui scientifique, il a été gentil, compréhensif vis à vis de mes responsabilités parentales, accueillant durant mes séjours en stage au sein du laboratoire. Il a toujours été disponible pour m'orienter et me conseiller au cours de l'élaboration de cette thèse. Je te remercie infiniment et j'espère que cette thèse est à la hauteur de tes espérances.

Je voudrais aussi témoigner ma profonde reconnaissance à Monsieur le professeur Pablo Gregori qui m'a beaucoup aidé, pour tous ces apports scientifiques et pour le temps qu'il a consacré à ma recherche. Son expérience dans le domaine et son esprit perfectionniste, ont significativement influencé cette thèse.

Je tiens à remercier chaleureusement Mme Houda El-bouhissi la présidente du jury pour son soutien moral et pour l'attention qu'elle a portée à ma soutenance de thèse.

Je remercie les membres de jury pour avoir accepté d'être les rapporteurs de ma thèse, pour le temps qu'ils ont consacré pour sa lecture, et pour l'attention avec laquelle ils l'ont lue et évaluée.

Je tiens à témoigner ma reconnaissance à madame Malika Yaici pour sa lecture minutieuse de ma thèse. Ces remarques ont amélioré la qualité de rédaction de la thèse.

Toute ma reconnaissance va également vers mes chers parents, mes sœurs et frères de m'avoir encouragée et toujours aidée, tout particulièrement ma mère qui me donne toujours la force de ne jamais baisser les bras. Je remercie également mon conjoint qui a droit lui aussi à toute mon appréciation, il était toujours à mes côtés, ma belle-famille, mes enfants Khalil et Ahmed qui me donnent le courage pour aller de l'avant.

Je remercie toute l'équipe de recherche AND de l'Institut FEMTO-ST à Belfort en France qui m'a accueillie plusieurs fois.

Je tiens à remercier particulièrement le chef de notre laboratoire LIMED Monsieur le professeur Hachem Slimani, Hayette avec laquelle j'ai partagé de bon moment, et je souhaite la réussite à tous les membres du laboratoire LIMED.

Table des matières

Table des matières i

Table des figures iv

Notations et symboles v

Liste des contributions vi

Introduction 1

1 Etat de l'art sur l'analyse statistique implicatif, (R)CHIC et la classification 13

1.1 Introduction . . . 14

1.2 L'Analyse Statistique Implicative (ASI) . . . 14

1.2.1 Origine, définition et méthodologie . . . 14

1.2.2 Modélisation mathématique de l'approche de base . . . 15

1.2.3 Extension de l'ASI . . . 16

1.2.4 Autres indices . . . 17

1.3 (R)CHIC . . . 19

1.3.1 Données/ Variables . . . 20

1.3.2 Graphe implicatif . . . 21

1.3.3 Arbre cohésif . . . 23

1.3.4 Arbre des similarités . . . 23

1.4 Domaines d'application de l'ASI . . . 24

1.5 Classification . . . 27

1.5.1 Présentation des Algorithmes de classification . . . 28

1.5.2 Études comparatifs des algorithmes de classification . . . 31

1.6 Conclusion . . . 32

2 Analyse des notes des étudiants par le logiciel CHIC 33

2.1 Introduction . . . 34

2.2 Étude des notes des étudiants en informatique de l'université de Bejaia . . . 35

2.2.1 Étude des notes des étudiants licence 2 durant les années scolaires 2010-2011, 2011-2012 ainsi que 2012-2013 . . . 37
2.2.2 Interprétation des résultats qui se répètent dans les trois générations . . . 40
2.2.3 Étude des notes des étudiants licence 3 durant les années scolaires 2010-2011, 2011-2012 ainsi que 2012-2013 . . . 41
2.2.4 Interprétation des résultats qui se répètent dans les trois générations . . . 43
2.3 Conclusion . . . 44

3 Ajout de la Confiance au graphe Implicatif 45
3.1 Introduction . . . 46
3.2 Motivations . . . 47
3.3 L'approche proposée et son application sur des données issues des échocardiographie de stress . . . 49
3.3.1 Résultats obtenus sans spécifier un seuil de confiance . . . 53
3.3.2 Résultats obtenus en utilisant un seuil de confiance égal à 80 . . . 54
3.3.3 Résultats obtenus en utilisant un seuil de confiance égale à 70 . . . 55
3.3.4 Résultats obtenus avec un seuil de confiance égal à 65 . . . 55
3.4 Conclusion . . . 56

4 Classification en utilisant les règles d'implication de l'ASI 57
4.1 Introduction . . . 58
4.2 Pourquoi l'ASI pour faire de la classification . . . 58
4.3 Description des ensembles de données . . . 60
4.4 Critères d'évaluation . . . 61
4.5 Approche 1 : Classification avec partitionnement de données . . . 63
4.5.1 Description de l'approche . . . 63
4.5.2 Exemple d'application de l'approche . . . 66
4.5.3 Expérimentation . . . 70
4.5.4 Evaluation . . . 71
4.6 Approche 2 : Classification sans partitionnement de données . . . 74
4.6.1 Description de l'approche . . . 74
4.6.2 Implémentation de l'approche . . . 76
4.6.3 Exemple d'application . . . 77
4.6.4 Résultats et discussions . . . 78
4.7 Conclusion . . . 82

Conclusion générale 83

Bibliographie 86

Table des figures

1 Diagramme de Venn pour la règle $a \rightarrow b$ 5

1.1 Représentation par les diagrammes d'Euler 15
1.2 Extrait d'un fichier contenant des données traitées par le logiciel (R)CHIC 20
1.3 Extrait du fichier transaction.out 21
1.4 Exemple d'un graphe implicatif 22
1.5 Exemple d'une boite de dialogue pour l'ajout ou la suppression d'une variable. . . . 22
1.6 Exemple d'un arbre cohésitif 23
1.7 Exemple d'un arbre de similarité 24

2.1 Exemple de graphe implicatif. 36
2.2 Graphe implicatif Licence 2 2010-2011. 38
2.3 Graphe implicatif Licence 2 2011-2012. 39
2.4 Graphe implicatif Licence 2 2012-2013. 40
2.5 Graphe implicatif Licence3 2010-2011. 42
2.6 Graphe implicatif Licence3 2011-2012. 42
2.7 Graphe implicatif Licence3 2012-2013. 43

3.1 Fenêtre permettant le choix de la valeur de confiance. 50
3.2 Extrait du jeu de données. 52
3.3 Extrait du fichier transaction.out correspondant au jeu de données. 52
3.4 Extrait du graphe d'implication avec l'affichage des valeurs de confiance. 53
3.5 Graphe implicatif avec un seuil de confiance égale à 80. 54
3.6 Graphe implicatif avec un seuil de confiance égale à 70. 55
3.7 Graphe implicatif avec un seuil de confiance égale à 65. 56

4.1 Extrait du jeu de données WBC partitionné en deux échantillons. 67
4.2 Extrait du fichier transaction.out. 67
4.3 Règles obtenus avec la variable V_9 en prémisse. 68
4.4 Exemple de liste de combinaison de variables. 68
4.5 La prédiction en utilisant chaque variable pour chaque individu. 68
4.6 Liste de règles obtenues. 69

4.7 Les combinaisons de règles correspondantes à la combinaison (V_6, V_2, V_1). 69
4.8 Simplification des règles de la Figure 4.7. 70
4.9 Extrait du jeu de donnée WBC. 77
4.10 Représentation graphique du Tableau 4.10. 81
11 Représentation par les diagrammes d'Euler 96

Notations et symboles

A	ASI	Analyse statistique implicative
C	CBA	Classification Based on Association rules.
	$CHIC$	Classification Hiérarchique Implicative et Cohésitive.
	CV	Cross-Validation.
	$CART$	Classification And Regression Tree.
D	DM	Data Mining.
E	ECD	Extraction des Connaissances a partir des Données.
K	KM	Knowledge Management.
M	ML	Machine Learning.
N	NB	Naïve Bayes.
R	RBF	Réseaux de neurones à Base radiale.
S	SVM	Support Vector Machines.
	SMO	Sequential Minimal Optimization.
U	UCI	University of California, Irvine.
W	WBC	Wisconsin Breast Cancer.
	$WDBC$	Wisconsin Diagnosis Breast Cancer.
	$WPBC$	Wisconsin Prognosis Breast.

Liste des contributions

Dans le cadre de cette thèse, nous avons réalisé les contributions scientifiques suivantes :

1. 2021 : Publication d'un papier dans le journal Mathematics 2021 (MDPI) Impact Factor 2.258. Intitulé de la publication « An Accurate and Easy to Interpret Binary Classifier Based on Association Rules Using Implication Intensity and Majority Vote »
2. 2019 : Participation au $10^{\text{ème}}$ Colloque International sur l'Analyse Statistique Implicative (ASI10) organisé en 2019 à Belfort en France. Intitulé de la communication : « Variabilité des arbres de similarité de l'Analyse Statistique Implicative sous indépendance statistique de variables binaires ».
3. 2017 : Participation au $9^{\text{ème}}$ Colloque International sur l'Analyse Statistique Implicative (ASI9) organisé en 2017 à Belfort en France. Intitulé de la communication :« Classification en utilisant les règles d'implication de L'ASI ».
4. 2015 : Participation au $8^{\text{ème}}$Colloque International sur l'Analyse Statistique Implicative (ASI8) organisé en 2015 en Tunisie. Intitulé de la communication :« Ajout de la confiance au graphe implicatif ».
5. 2014 : Participation au $4^{\text{ème}}$ International Symposium ISKO-Maghreb 2014 on Concepts and Tools for Knowledge Management (KM) organisé en 2014 en Algérie. Intitulé de la communication « Analyse of Bejaia University Computer Science student's marks through the CHIC software and Statistics Implicative Analysis ».

Introduction

L'augmentation des services numériques et des capacités de sauvegarde fait exploser le volume de données stockées sous forme numérique dans le monde. L'ECD (Extraction des connaissances dans les données) consiste à mettre en évidence des connaissances nouvelles, valides, et potentiellement utiles dans de grandes bases de données [WFM92]. En ECD, une des principales méthodes produisant des connaissances sous forme de règles est l'extraction de règles d'association. La plupart des algorithmes permettant d'extraire des règles d'association sont fondés sur la confiance et le support. Ils apportent des solutions au problème de l'extraction de règles, mais ils produisent une trop grande masse de règles, sélectionnant certaines règles sans intérêt et ignorant des règles intéressantes. Pour remédier à ce problème, de nombreuses mesures d'intérêt ont été développées afin d'évaluer les règles selon différents points de vue. Ces mesures tentent de représenter les liaisons associées entres variables, et dépendent du type de variables traités (binaire, multimodale), de la relation (symétrique ou asymétriques) et de l'association (exprimant une corrélation ou causalité) entre ces variables. Trois types de liaisons sémantiques implicatives distinctes pouvant être découvertes dans les données : Certaines expriment pour l'utilisateur une description, d'autres une causalité ou une corrélation. Les mesures d'intérêt sont nombreuses, plusieurs d'entre elles corrigent la principale critique faite à la confiance, mais elles héritent par construction de différentes caractéristiques de cette dernière. Cependant, le peu de mesures qui ne se réduisent pas à la confiance sont symétriques. Elles ne permettent pas l'extension aux méta-règles, elles tendent à être peu discriminantes quand la taille des phénomènes étudiés sont grands [EP96] et écrasent les cas rares, appelés « pépites de connaissances » dans les recherches d'Yves Kodratoff. En terme de qualité, les relations causales (asymétriques) sont considérées plus intéressantes que les relations de corrélation (symétriques). Une relation causale permet, d'acquérir une capacité prédictive et offre la possibilité de mieux maîtriser l'enchaînement des phénomènes et d'ordonner des variables en séquence implicative. Pour atteindre cet objectif il convient de faire appel à des indices de liaisons qui ne sont pas symétriques comme ceux de l'Analyse Statistique Implicative (ASI).

L'ASI rapproche la règle de l'implication logique, et prend en considération la nature des règles d'association transactionnelles. La mesure de base conçue initialement pour cette méthode est appelée l'intensité d'implication. Elle prend en considération la non satisfaction de l'implication et elle est dissymétrique. Cette mesure a impliqué un grand nombre de travaux de recherche

et d'application à travers une interface graphique, le logiciel CHIC (Classification hiérarchique Implicative et Cohésitive).

Notre but dans cette thèse est de :

1. Montrer l'apport de l'ASI et l'efficacité du logiciel (R)CHIC [1].
2. Améliorer et créer certaines fonctions permettant de calculer les méthodes de l'ASI dans R.
3. Rajouter d'autres fonctionnalités au logiciel RCHIC que CHIC ne possède pas.
4. Obtenir un classificateur ayant du sens et facile à interpréter avec les règles d'implication de l'ASI.

Règles d'association

De nombreuses théories de représentation de la connaissance sont fondées sur les règles [HHNT86]. D'une manière générale, les règles sont des propositions de la forme "si prémisse alors conclusion", notées prémisse $\rightarrow$ conclusion. Ces règles signifient que si un enregistrement de la table vérifie la prémisse, alors il vérifie sûrement également la conclusion. Elles ont l'avantage de représenter les connaissances de manière explicite. Les règles sont dotées de plusieurs mesures de qualité (ou d'intérêt). Les plus utilisées sont la confiance, le support et le lift.

Le support d'une règle [AS+94] est la proportion de transactions qui réalisent à la fois la prémisse et la conclusion de la règle. la confiance [AIS93] est la proportion de transactions qui réalisent la conclusion, parmi celles qui réalisent la prémisse, c'est-à-dire la fréquence relative conditionnelle de la conclusion sachant la prémisse. Le lift [BMS97b] est le quotient du nombre d'exemples observés par celui attendu sous l'hypothèse d'indépendance de la prémisse et de la conclusion. C'est la proportion entre la confiance de la règle et le support. Ces mesures sont formalisées de la manière suivante :

Soient n(A) et n(B) les nombres de transactions qui réalisent respectivement les items de A et de B, n(AB) le nombre de celles qui réalisent à la fois A et B.

$$Supp(A \rightarrow B) = P(AB) = n(AB)/n. \quad \text{(Support)}$$

$$Conf(A \rightarrow B) = P(AB)/P(A) = n(AB)/n(A) \quad \text{(Confiance)}$$

$$L(A \rightarrow B) = P(AB)/P(A).P(B) \quad \text{(Lift)}$$

Une des principales méthodes produisant des connaissances sous forme de règles est l'extraction de règles d'association, introduite par Agrawal, Imielienski et Swami [AIS93]. Depuis, de nombreux algorithmes ont été développés pour extraire efficacement des règles d'association (voir [HGG00] pour une synthèse). La plupart de ces algorithmes valident les règles avec les deux mesures : le support et la confiance. Ces algorithmes parcourent les enregistrements pour rechercher ceux dont

1. RCHIC est la version écrite en R du logiciel CHIC

le support dépasse un certain seuil (minsupp), pour en déduire les règles d'association dont la confiance dépasse le seuil de confiance (minconf).

Avantage et inconvénient de l'approche support-confiance

L'approche support-confiance présente un grand intérêt comme critères d'extraction vu l'importance du support et de la confiance. Le sens concret des valeurs du support et de la confiance est parfaitement assimilable par l'utilisateur non spécialiste. L'approche est clairement définie et déterministe, au sens où tous les algorithmes doivent découvrir les mêmes règles d'association, celles qui vérifient les conditions de support et de confiance. Mais ils produisent une trop grande masse de règles, sélectionnant certaines règles sans intérêt et ignorant des règles intéressantes.

En outre, la condition de support qui est le moteur même du processus d'extraction écarte les règles ayant un petit support alors que certaines peuvent avoir une très forte confiance et présenter un réel intérêt. Si l'on baisse le seuil de support pour remédier à cet inconvénient, les règles générées sont nombreuses et les algorithmes d'extraction sont débordés. De plus, les règles sélectionnées par cette approche deviendront triviales et non intéressantes dans le cas où la conclusion de la règle est très fréquente dans l'échantillon, car une grande confiance n'implique pas nécessairement que la permisse induit un effet positif sur la conclusion. Les seules conditions de support et de confiance ne suffisent pas à assurer le réel intérêt d'une règle. Pour remédier à ce problème, de nombreuses mesures d'intérêt ont été développées afin d'évaluer les règles selon différents points de vue [Fre98] [BMS97a][PS91].

Ces mesures permettent à l'utilisateur d'identifier et de rejeter les règles de faible qualité, mais aussi d'ordonner les règles acceptables des meilleures aux plus mauvaises [Bla05]. Ces mesures sont des mesures de la liaison associées entres variables et dépendent du type de variables traitées (binaire, multimodale), de la relation (symétrique ou asymétrique) et de l'association (expriment une corrélation ou causalité) entre ces variables.

Nature des liaisons entre variables

Nous distinguons deux types de liaison entre les variables : les liaisons symétriques et asymétriques. Les mesures symétriques, évaluent de la même façon les règles $a \rightarrow b$ et $b \rightarrow a$, et les liaisons asymétriques évaluent ces deux dernières d'une manière différente [LT04]. Pour bien illustrer la différence entre les deux types de liaison nous prenons l'exemple de variable qualitatif et binaire. Les liaisons entre les variables qualitatives nominales multimodales sont symétriques, car les différentes modalités d'une variable sont traitées de la même manière et cette liaison n'est pas affectée par la permutation des modalités d'une des variables. Elles sont plutôt utilisées en classification où toutes les modalités de la variable classe doivent être exprimées. Étant données la règle $(a \rightarrow b)$, ou b représente la variable classe, nous supposons que les différentes modalités de

la variable classe sont : (petit, moyen et grand). La variable a est partitionnée en échantillons afin d'affecter chaque échantillon à la modalité correspondante. Par exemple ($a1 \rightarrow petit$, $a2 \rightarrow moyen$ et $a3 \rightarrow grand$). Un individu classifié comme petit appartient belle et bien à l'échantillon 1 de la variable a, donc les règles a $\rightarrow$ b et b $\rightarrow$ a sont traités de la même manière. Les liaisons entre variables binaires sont asymétriques car pour deux variables binaires a et b, si l'on considère b est la négation de a dans le cas booléen, alors la liaison entre a et $\bar{a}$ est différente de la liaison entre $\bar{a}$ et a. Cette distinction entre les deux types de liaisons est importante car elle justifie que les mesures de liaison entre variables multimodales sont en général moins adaptées à l'évaluation des liaisons entre variables binaires.

Sémantique des liaisons entre variables

Kodratoff [Kod00] fait la distinction entre deux types de liaisons sémantiques implicatives pouvant être découvertes dans les données. Certaines liaisons notées a $\rightarrow$ b expriment pour l'utilisateur une description, comme "les corbeaux sont noirs" (corbeau $\rightarrow$ noir). Les liaisons qui n'expriment pas de description peuvent exprimer soit une causalité ou une corrélation. Le concept de causalité ou de nécessité de a sur b signifie que b se produit quant ou dès que a se produit. Par exemple, la crise économique produit du chômage. Dans ce cas, il y a une relation de cause entre la crise et le chômage. Le concept de corrélation (concomitance), vise à déterminer la réalisation simultanée de deux événements respectifs à deux variables a et b . Par exemple, l'achat du lait et du pain au supermarché sont deux faits concomitants. On ne peut pas dire que c'est à cause de l'un que l'autre produit a été acheté (le lait et le pain sont achetés ensemble).

En terme de qualité, les relations causales sont considérées plus intéressantes que les relations de corrélation. Une des raisons, est l'usage que l'on peut faire de la connaissance des relations. En effet, une relation causale permet, d'acquérir une capacité prédictive. Cette capacité offre la possibilité de mieux maîtriser l'enchaînement des phénomènes, d'où la possibilité d'anticiper sur des enchaînements pour mieux servir les besoins de l'utilisateur. Par exemple, comprendre les causalités entre les phénomènes météorologiques permet d'éviter certaines catastrophes naturelles.

Un lien symétrique ne peut pas être un lien de causalité. Il est impossible d'ordonner des corrélations entre variables en séquence implicative.

Règle, implication, équivalence et quasi-implication

Les liaisons qui expriment des descriptions correspondent précisément à la définition des règles. Les liaisons qui expriment des corrélations correspondent à la définition de l'équivalence. Tandis que les liaisons qui expriment des causalités correspondent à la définition de la quasi-implication.

Une règle est un couple de variables booléennes (a, b) noté $a \rightarrow b$, tel que, a et b appartenant respectivement aux ensembles A et B. a est la prémisse de la règle et b sa conclusion. Elle traduit

la tendance de b à être vrai quand a est vrai, et peut se lire de la manière suivante : "si un individu vérifie a alors il vérifie sûrement b". Les exemples d'une règle sont les individus de $A \cap B$, c'est-à-dire ceux qui vérifient la prémisse et la conclusion, tandis que les contre-exemples sont les individus de $A \cap \bar{B}$, ceux qui vérifient la prémisse mais pas la conclusion (Figure 1). Une règle est d'autant meilleure qu'elle admet beaucoup d'exemples et peu de contre-exemples. Par conséquent, à partir de deux variables a et b il est possible de construire plusieurs règles déférentes. Pour une règle $a \rightarrow b$, $a \rightarrow \bar{b}$ est la règle contraire, $b \rightarrow a$ est la règle réciproque, et $\bar{b} \rightarrow \bar{a}$ est la règle contraposée [Bla05].

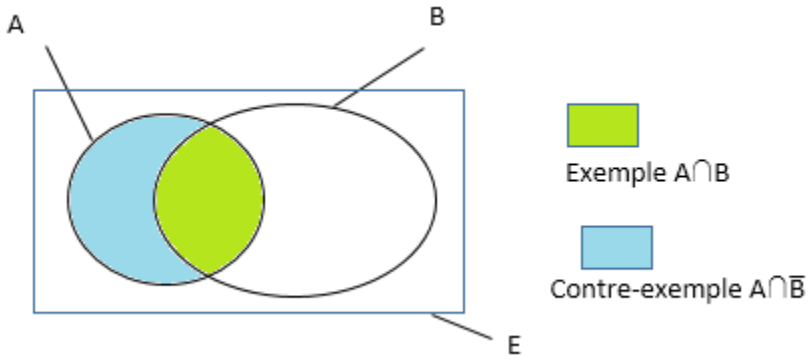

FIGURE 1 – Diagramme de Venn pour la règle $a \rightarrow b$

Dans le Tableau 1 (a), (b) , (c) et (d) nous avons représenté respectivement la table de contingence de la règle $a \rightarrow b$, la table de vérité de l'implication logique $a \Rightarrow b$, la table de vérité de l'équivalence $a \Leftrightarrow b$ et la table de contingence de la quasi-implication. Les cas ($a = 1$ et $b = 1$) et ($a = 1$ et $b = 0$) possèdent le même rôle pour la règle et l'implication, les premiers vérifient la règle et l'implication, et les seconds les contredisent. Les cas ($a = 0$ et $b = 1$) et ($a = 0$ et $b = 0$) ne jouent pas le même rôle pour $a \rightarrow b$ et $a \Rightarrow b$: ils vérifient l'implication mais ne sont pas des exemples pour la règle. Pour $a \rightarrow b$, ces cas n'ont pas de rôle défini [LT04]. Une règle traduit uniquement la tendance de la conclusion à être vraie quand la prémisse est vraie.

Les liaisons qui expriment des descriptions doivent être infirmées chaque fois que (a = 1 et b = 0) est observé, et confirmées chaque fois que (a = 1 et b = 1) est observé. Comme le suggère le paradoxe de Hempel9. Les liaisons notées $a \Rightarrow b$ expriment pour l'utilisateur une causalité, doivent être infirmées chaque fois que ($a = 1$ et $b = 0$) est observé, et confirmées chaque fois que ($a = 1$ et $b = 1$) ou ($a = 0$ et $b = 0$) est observé.

En revanche, une liaison qui exprime une causalité n'est pas une règle au sens strict : en considérant les cas ($a = 0$ et $b = 0$) comme des exemples, elle traduit à la fois la règle $a \rightarrow b$ et la règle contraposée $\bar{a} \rightarrow \bar{b}$. Ce que nous appelons une quasi-implication. Les liaisons exprimant des causalités approximent donc mieux que les règles d'implication logique.

Les liaisons qui expriment des corrélations correspondent à la définition de l'équivalence ($a \Leftrightarrow b$) cela revient à dire que a et b ont les même valeur de vérité : a et b sont soit tous les deux vrais,

a \ **b**	1	0
1	n_{ab} (exemple)	$n_{a\bar{b}}$ (contre-exemple)
0	$n_{\bar{a}b}$	$n_{\bar{a}\bar{b}}$

(a) Table de contingence de la règle $a \rightarrow b$

a \ **b**	1	0
1	1	0
0	0	1

(b) Table de vérité de l'équivalence logique $a \Leftrightarrow b$

a \ **b**	1	0
1	1	0
0	1	1

(c) Table de vérité de l'implication logique $a \Rightarrow b$

a \ **b**	1	0
1	n_{ab} (exemple)	$n_{a\bar{b}}$ (contre-exemple)
0	$n_{\bar{a}b}$	$n_{\bar{a}\bar{b}}$ (exemple)

(d) Table de contingence de la quasi-implication $a \Rightarrow b$

TABLEAU 1 – Comparaison entre : règle, implication, équivalence et quasi-implication

soit tous les deux faux.

Si l'utilisateur est intéressé par des liaisons exprimant des descriptions, alors il est préférable d'employer des indices de règle (au sens strict) et d'éviter les indices de quasi-implication. Au contraire, si l'utilisateur est intéressé par des liaisons exprimant des causalités, alors il vaut mieux employer des indices de quasi-implication et éviter les indices de règle, qui peuvent attribuer à une règle et sa contraposée des valeurs contradictoires [Bla05].

Critères d'appréciation d'une mesure d'intérêt

Les mesures d'intérêts, comme leurs noms l'indiquent, servent à évaluer ou à valider les règles d'association. De nombreuses mesures d'intérêts ont été développées pour compléter le support et la confiance qui, utilisés seuls, ne permettent d'évaluer que certains aspects de la qualité des règles. Plusieurs auteurs se sont intéressés aux propriétés qu'une bonne mesure doit vérifier [PS91] [TKS04] [LT04] [GCB+04]. Plusieurs travaux présentent des études comparatives formelles ou expérimentales des indices tel que : [JBA99] [TKS04]. Vaillant et al. [VLL04] quant à eux réalisent une étude expérimentale approfondie en comparant une vingtaine d'indices sur différentes bases de règles. Ils mettent en évidence trois groupes d'indices similaires globalement stables sur les bases considérées. Gras et al., Lallich et Teytaud proposent plusieurs critères d'évaluation des indices et cela dans le cadre de leur participation au groupe de travail GafoQualité de l'action spécifique STIC [BSGG04], Nous allons résumer les principaux critères abordés par ces travaux :

- La mesure a-t-elle un sens concret qui soit parlant pour l'utilisateur ? certaines mesures sont faciles à interpréter comme (le support, la confiance et le lift). Par exemple un lift > 1 pour la règle $A \rightarrow B$ signifie que la réalisation de A augmente les chances de la réalisation de B. tandis que d'autres mesures tel que la J-mesure et la forme entropique de l'intensité d'implication sont moins aisée à interpréter.
- Une mesure doit distinguer la règle $A \rightarrow B$ de la règle contraire $A \rightarrow \bar{B}$, les exemples de l'une étant les contre-exemples de l'autre. Le coefficient de corrélation [Pea96] prend en considération la différence entre ces deux règles, contrairement à la mesure du $\chi 2$ [Ler] et à la J-mesure[GS88].
- On préférera les mesures dissymétriques qui respectent la nature des règles d'association transactionnelles. Les mesures symétriques comme le support, la mesure de Piatetsky-Shapiro [PS91], le lift ou le coefficient de corrélation et ses dérivés, évaluent de la même façon les règles $A \rightarrow B$ et $B \rightarrow A$; alors que celles-ci ont les mêmes exemples mais pas les mêmes contre-exemples.
- Une mesure ne doit pas évaluer de la même façon les règles $A \rightarrow B$ et $\bar{B} \rightarrow \bar{A}$ [Kod00]. En effet, les deux règles ont les mêmes contre-exemples, mais elles n'ont pas les mêmes exemples. La prise en compte de la contraposée, ainsi dans l'intensité d'implication entropique [GKCG01], rapproche la règle de l'implication logique.

— Approche descriptive ou bien statistique : On appelle mesure descriptive une mesure qui ne change pas en cas de dilatation des données. Sinon la mesure est dite statistique. Une mesure statistique suppose que l'on ait un modèle aléatoire et une hypothèse $H0$ exprimant l'indépendance de A et B.

— Sens de variation de la mesure : Moins une règle a de contre-exemples, plus elle est intéressante. Une mesure doit donc être décroissante en fonction du nombre de contre-exemples. Une règle est d'autant plus intéressante que son nombre de contre-exemples (resp. Exemples) est exceptionnellement bas (resp. haut) sous l'hypothèse d'indépendance de A et B.

— Pouvoir discriminant : Les mesures issues d'une approche statistique ont tendance à perdre leur pouvoir discriminant lorsque le nombre de transactions n est grand. Plusieurs solutions ont été conçus pour corriger cette perte de pouvoir discriminant. Dans le cas de l'intensité d'implication, Gras, Kuntz, Couturier et Guillet (2001)[GKCG01] suggèrent la prise en compte d'un indice d'inclusion pour corriger la perte de discrimination.

— Classement induit par une mesure : Deux mesures m et m1 classent dans le même ordre les règles extraites d'une base de données si et seulement si pour tout couple de règles extraites de la base $m(A \to B) > m(C \to D) \Leftrightarrow m1(A \to B) > m1(C \to D)$ On définit ainsi une relation d'équivalence sur l'ensemble des mesures possibles.

Les mesures d'intérêt

Les mesures d'intérêt sont nombreuses mais elles tendent généralement vers la confiance soit par normalisation en utilisant des transformations affines [2] ou selon le classement induit par ces mesures. Les principales mesures (Sebag-Schoenauer [SS88], Loevinger [Loe47], Corrélation, Lift, Multiplicateur de cote [LT04], conviction [LT04]), corrigent la principale critique faite à la confiance, mais elles héritent par construction de différentes caractéristiques de la confiance [Gui02]. A titre d'exemple la mesure de Sebag-Schoenauer est classée comme la confiance, puisqu'elle s'écrit comme une transformation monotone croissante de la confiance : $Seb(a \to b) = (Conf(a \to b)/1 - Conf(a \to b))$. Le lift d'une règle est la proportion entre la confiance de la règle et le support de la conclusion. Il peut être interprété comme l'effet de la prémisse sur la conclusion. On peut avoir des règles avec un niveau de confiance élevé (ce qui donne une bonne capacité de prédiction), mais un $lift < 1$, signifie que la prémisse réduit les chances initiales de la conclusion. Donc le lift corrige la principale critique faite à la confiance. Mais le lift s'exprime à partir des seuls exemples, il est symétrique. C'est aussi une mesure de qualité descriptive, donc sujette à une variabilité naturelle dans le cadre du plan d'échantillonnage.

Parmi les mesures qui ne se réduisent pas à une transformée affine de la confiance, on citera

2. Les transformations affines sont utilisées pour placer des objets dans l'espace, ou pour effectuer des changements de repère.

notamment le $\chi 2$ et l'indice d'implication [LA02]. Ces deux mesures prennent en compte les contre-exemples ($a \rightarrow \bar{b}$). Le $\chi 2$ est un test statistique utilisé pour tester l'indépendance conditionnelle entre les propriétés a et b utilisées notamment par [BMS97a], cependant il est symétrique. Ces mesures ne permettent pas l'extension au méta règle du type $a \rightarrow (b \rightarrow c)$ et elles tendent à être peu discriminantes quand la taille des phénomènes étudiés sont grands [EP96].

L'Apport de l'ASI

Vu la nécessité de compléter le support et la confiance par d'autres mesures d'intérêt, et dans le but de palier aux limites d'autres mesures, Régis et all [GA96] ont proposé une mesure qui rapproche la règle de l'implication logique. Cette dernière prend en considération la nature des règles d'association transactionnelles : "si tels articles (a) sont dans le panier, alors le plus souvent tels autres (b) y sont, car dans le cas de l'implication logique l'égalité stricte est requise, mais elle ne l'est pas au sens des règles d'association. Cette mesure quantifie l'invraisemblance de la petitesse du nombre de contre-exemples $n_{a\bar{b}}$ eu égard à l'hypothèse d'indépendance entre a et b. Elle prend en considération la non satisfaction de l'implication (qui porte sur les contre exemples) et elle est dissymétrique. Cet indice est appelé l'intensité d'implication. Ce dernier est un indice de quasi-implication développé par Gras [GA96] et qui est au fondement d'une méthode d'analyse exploratoire de données nommée analyse statistique implicative (ASI)[GKB01].

L'Analyse statistique implicative est une méthode originale d'analyse de données à la recherche de causalités, elle offre une possibilité d'ordonner des variables en séquence implicative, ce qui est impossible avec les mesures d'intérêts symétriques. Elle est basée sur la mesure de l'Intensité d'implication. Comme toutes les mesures de significativité statistique, cet indice est peu discriminant quand les cardinaux étudiés sont grands (voir chapitre 1). Pour résoudre ce problème, Gras et al. ont proposé dans [GKB01] de moduler les valeurs de l'intensité d'implication par un indice de quasi-implication descriptif fondé sur l'entropie de Shannon : l'indice d'inclusion. Le nouvel indice ainsi formé s'appelle intensité d'implication entropique.

Ces mesures d'intérêts ont impliqué un grand nombre de travaux de recherche et d'applications à travers une interface graphique, le logiciel CHIC (Classification Hiérarchique Implicative et Cohésitive), qui permet une utilisation facile de diverses techniques en ASI pour un large éventail d'utilisation. Ce logiciel a été conçu il y a 20 ans et il n'est plus maintenu. Par ailleurs, le logiciel R [HCHB11] regroupe de très nombreuses méthodes à disposition de la communauté scientifique, les fonctionnalités de CHIC ont été réécrites avec le logiciel R.

Contributions de la thèse

Les contributions de la thèse se déclinent en cinq thèmes.

Tout d'abord, nous avons effectué un prérequis sur l'évaluation de l'intérêt des règles d'association pour assister l'utilisateur à comprendre l'apport de l'ASI. Nous avons montré que les seules conditions de support et de confiance ne suffisent pas à assurer le réel intérêt d'une règle et que d'autres mesures qui prennent en considération la nature, le type et la sémantique de la relation entre variables sont nécessaires. Ensuite nous avons fait une vue d'ensemble de la théorie ASI qui vient non pas pour s'opposer à la pratique traditionnelle mais pour enrichir la qualité des règles. Nous avons présenté différents travaux réalisés avec l'ASI dans différent domaines, et cela pour inciter les administrations à utiliser l'ASI. Par la suite nous avons présenté le logiciel conçu pour l'ASI (R)CHIC et quelques méthodes de classification en Data Mining.

La deuxième contribution porte sur l'étude des notes des étudiants en utilisant le logiciel CHIC, dont le but initial est de se familiariser avec le logiciel et de montrer l'utilité de cette théorie. Cette étude permet de fournir une formation scientifique utile pour les futures étudiants. A l'aide du graphe implicatif (voir section 1.3.2 page 21) nous avons établit les liens entre les matières étudiées, ce qui peut aider les étudiants à identifier les matières maitrisées qui sont nécessaires pour maitriser d'autres matières, ceci peut les aider dans leur orientation future. Cette étude offre aux administrations une idée sur le bon déroulement des enseignements fournis aux étudiants et permet d'identifier les raisons des échecs des étudiants.

La troisième contribution consiste à réaliser le premier objectif de la thèse qui consiste à améliorer et à créer certaines fonctions permettant de calculer les méthodes de l'ASI dans R. Il s'agit de poursuivre des travaux entrepris à ce sujet. De nombreux travaux existent sur l'ASI mais en dehors de R. Le fait d'intégrer les travaux majeurs permettra à la communauté de bénéficier des avantages de R (simplicité, portabilité, reconnaissance) et permettra également d'améliorer la diffusion de l'ASI. Nous avons en premier lieu rendu les noeuds du graphe d'implication mobile, cela à permet à l'utilisateur d'organiser et de redimensionner les graphes à sa guise.

Par la suite d'autres fonctionnalités ont été rajouté au logiciel RCHIC que CHIC ne possède pas. Nous avons ajouté une nouvelle option pour la sélection des règles dans le graphe implicatif qui est le couplage de l'indice d'implication avec la confiance afin d'améliorer, de faciliter la recherche de règle avec la méthode d'analyse statistique implicative et aider les experts et utilisateurs à utiliser le graphe d'implication. Cela est mis en œuvre en ajoutant la valeur de la confiance pour chaque règle dans le graphe implicatif, ce qui permet de distinguer le niveau d'importance de chaque règle. Afin de rendre le graphe plus lisible l'utilisateur utilise un seuil pour le choix de la valeur de confiance. L'approche a été testée sur différents ensembles de données.

Dans la quatrième contribution nous avons proposé une nouvelle méthode de classification basée sur les règles d'implication de l'ASI. Notre préoccupation majeure est d'avoir des règles qui ont du sens, facile à interpréter et ayant une erreur de prédiction faible. Cette méthode est différente des méthodes de classifications basées sur les règles d'association (CBA : class based association) du fait qu'elles utilisent une mesure de qualité plutôt nouvelle, l'implifiance (voir sec-

tion 1.2.4.4 page 18). Dans cette méthode, le jeu de données est partitionné en un nombre fixe d'intervalles et les données sont analysées afin de trouver des relations entre l'appartenance d'un élément à un échantillon donné d'une variable et son appartenance à une classe donnée. Ces relations sont formalisées sous forme de règles d'implications. Toutes les règles importantes possibles seront sélectionnées. Pour prédire l'état des éléments nous utilisons une combinaison de règles. Cette dernière est réalisée en combinant les variables contribuant à la formation des règles déjà sélectionnées. L'utilisateur choisit le nombre de règle maximum que peu contenir une combinaison noté nvote. Nous sélectionnons toutes les combinaisons de nvote variables, nous calculons l'erreur de prédiction pour chaque combinaison de variables existantes, ensuite nous sélectionnons la combinaison dont l'erreur de prédiction est la plus faible. Les règles utilisées pour faire de la prédiction correspondent à celle déduite à partir de la combinaison ayant la plus petite erreur de prédiction. Les règles obtenues seront simplifiées pour avoir moins de règles avec moins de conjonctions à présenter à l'utilisateur pour faire de la classification. Pour illustrer cette proposition, notre approche a été expérimentées sur des ensembles de données sur le cancer de sein en libre accès, et les résultats obtenus ont été comparés à quatre autres algorithmes de classification bien établis.

La dernière contribution consiste à proposer un nouveau classificateur basé sur les règles d'implication de l'ASI, qui doit vérifier encore plus de contraintes que l'approche précédente. En plus de sa facilité d'interprétation et de sa précision nous avons fait de sorte que la comparaison et l'évaluation de notre approche par rapport aux autres approches soit équitable. Pour cela nous avons appliqué la validation croisée "cross-validation" à notre classificateur ainsi qu'aux autres méthodes. Cette approche offre plus d'avantages car : premièrement l'approche de classification est plus facile à utiliser, car un simple utilisateur puisse : installer le package R, recopier le code, trouver les mêmes résultats publiés, et plus encore, changer les paramètres librement et trouver d'autres solutions car nous avons mis de sorte à exécuter l'approche en dehors de Rchic. Nous avons également reproduit les mesures d'intérêt qui nous intéresse, tel que l'implifiance (voir section 1.2.4.4 page 18). Dans le but d'avoir des résultats plus précis nous avons raisonné d'une manière différente pour faire les prédictions. Dans cette nouvelle approche, un individu (ligne) est classé directement en fonction des valeurs des variables significatives sélectionnées (colonne). Contrairement a la première approche où chaque variable est partitionné, donc les prédictions se font selon l'échantillon qui contient la variable de l'individu dont on veut prédire. Nous avons également reproduit les résultats des algorithmes de DATA Mining, avec le package RWeka, ce qui nous a permis d'exécuter tous les algorithmes dans les mêmes conditions. Nous avons également testé notre approche par de nouveaux jeux de données en plus des jeux de données sur le cancer de sein. Les résultats obtenus ont été comparés aux résultats obtenus des autres algorithmes de classification.

Organisation de la thèse

Cette thèse est constituée de cinq chapitres.

Dans le premier chapitre nous introduisons les concepts qu'un utilisateur ASI doit acquérir pour assimiler l'apport de l'ASI. Nous mettons le point sur les limites de l'approche support confiance, sur la nature et la sémantique des liaisons entre variables. Nous expliquons, la différence entre une règle d'association, une implication, une corrélation et une quasi-implication, les différentes mesures d'intérêt ainsi que l'apport de l'ASI sur l'enrichissement de ces mesures.

Dans le chapitre 2 nous réalisons un état de l'art sur la théorie ASI. Nous présentons la modélisation mathématique de l'approche de base et des principales mesures d'intérêts (l'indice entropique, l'indice d'inclusion et l'implifiance) sur lesquelles se base la méthode ASI. Nous décrivons le logiciel (R)CHIC en présentant ces différentes fonctionnalités, exposons les domaines d'applications de l'ASI et enfin, une présentation de la classification en Data Mining et une description de quelques algorithmes de classification.

Dans le troisième chapitre nous offrons aux utilisateurs un nouveaux mode pour le calcul du graphe d'implication, le mode intensité d'implication + confiance. Nous organisons le chapitre comme suit : nous calculons l'intensité d'implication et la confiance pour des données expérimentales et à travers les résultats obtenus, nous montrons comment l'expert peut utiliser ces deux informations ensemble pour extraire les règles qui l'intéressent le plus. Ensuite nous appliquons l'approche a des données collectées à partir de L'UCI, puis nous présentons les expérimentations et les différents résultats obtenus avec des interprétations.

Dans le dernier chapitre, nous présentons deux approches de classification avec l'ASI. Nous justifions notre choix d'utilisation de l'ASI pour faire de la classification puis nous donnons une brève description des jeux de données en libre accès utilisés pour l'apprentissage et pour le test. Nous décrivons les critères d'évaluation utilisés pour évaluer nos classificateurs et présentons les deux méthodes de classifications ainsi que la logique de chaque classificateur. Nous divisons le chapitre en deux parties :

— Nous consacrons la partie 1 pour décrire le premier classificateur. Nous exposons sa logique ainsi qu'un exemple d'application détaillé, et aussi expérimentons et évaluons ce classificateur avec différents jeux de données, en utilisant les outils du Data Mining.

— Nous réservons la partie 2 au deuxième classificateur. L'introduction expose les apports de ce classificateur par rapport au premier. Par la suite, nous décrivons les étapes suivies pour réaliser le classificateur. Nous présentons un exemple d'application pour voir la sortie de notre classificateur. Puis nous exposons et discutons les résultats d'exécution de notre approche par rapport aux autres approches de Data Mining. Finalement nous décrivons les étapes à suivre pour lancer le classificateur.

Chapitre 1

Etat de l'art sur l'analyse statistique implicatif, (R)CHIC et la classification

Sommaire

1.1 Introduction 14
1.2 L'Analyse Statistique Implicative (ASI) 14
1.2.1 Origine, définition et méthodologie 14
1.2.2 Modélisation mathématique de l'approche de base 15
1.2.3 Extension de l'ASI 16
1.2.4 Autres indices 17
1.3 (R)CHIC 19
1.3.1 Données/ Variables 20
1.3.2 Graphe implicatif 21
1.3.3 Arbre cohésif 23
1.3.4 Arbre des similarités 23
1.4 Domaines d'application de l'ASI 24
1.5 Classification 27
1.5.1 Présentation des Algorithmes de classification 28
1.5.2 Études comparatifs des algorithmes de classification 31
1.6 Conclusion 32

1.1 Introduction

Dans ce chapitre nous présentons un état de l'art sur l'analyse statistique implicative (ASI) et les différents concepts abordés dans les contributions proposées dans cette thèse. Nous effectuons tout d'abord une vue d'ensemble de la théorie ASI. Nous présentons le logiciel (R)CHIC (Classification Hiérarchique Implicative et Cohésitive) conçu pour implémenter les différentes fonctionnalités de cette analyse de données (ASI). Nous présentons également la classification en Data Mining ainsi que certains algorithmes de classification.

1.2 L'Analyse Statistique Implicative (ASI)

1.2.1 Origine, définition et méthodologie

L'ASI, à l'origine développée par Gras et ses collaborateurs [Gra79], est apparue suite aux difficultés rencontrées pour évaluer le niveau des élèves dans un test de mathématiques. Régis a enseigné les mathématiques à tous les niveaux d'enseignement dans plusieurs pays. Il a rencontré des difficultés d'apprentissage avec tous les niveaux, à l'école primaire, collège, lycée et également dans les niveaux supérieurs. Les difficultés rencontrées par Régis se résume dans l'évaluation du cas suivant. Certains élèves réussissent des tests jugés difficiles, tout en échouant à d'autres plus facile dans des contextes semblables. Ce problème a été formalisé avec des quasi-implications de la manière suivante : « la réussite à un item jugé difficile s'accompagne généralement de la réussite à un item plus facile ». L'évaluation et la structuration de telles relations implicatives entre situations didactiques sont les problèmes génériques à l'origine du développement de l'Analyse Statistique Implicative. Les premiers pas était en 1979 et continue toujours à se développer jusqu'à nos jours par Régis, ses collaborateurs et plusieurs d'autres chercheurs [Gra79].

Plusieurs autres phénomènes dans la vie peuvent être représentés par des règles de quasi implication. La quasi implication est représentée par la relation « si a alors généralement b », où a et b sont deux variables booléennes appartenant respectivement aux sous ensembles A et B de l'ensemble E. Ces règles peuvent être confrontées par un nombre important de réussite mais elles peuvent toujours avoir des contre-exemples. La théorie ASI prend en considération ces cas particuliers et n'abandonne pas une règle dès l'apparition d'un seul contre-exemple. Ce raisonnement a été montré par Laurent Fleury [Fle96]. L'ASI s'intéresse à ce genre de règles exprimant une causalité, ces règles sont asymétriques [GKG15]. La stratégie utilisée dans l'ASI consiste à prendre plutôt en considération la non-satisfaction de l'implication « si a alors b » qui, apparaît dès lors que a étant vrai, b est faux. Ce sont donc les contre-exemples sur lesquels vont porter les mesures de qualité de ces règles. D'où la définition de l'ASI donnée par Gras et al.

« L'Analyse Statistique Implicative a été définit comme étant « [...] un champ théorique centré sur le concept d'implication statistique ou plus précisément sur le concept de quasi-implication

pour le distinguer de celui d'implication logique des domaines de la logique et des mathématiques » [GRG09]. Gras et ses collaborateurs, ont pris justement ce quasi auquel ils ont donné un sens et une mesure ; initialement un indice, compris entre 0 et 1 a été établit, capable de rendre compte de l'écart entre ce qui était attendu et ce qui était effectivement observé. Cette mesure est relativisée par le nombre de données vérifiant respectivement a et non b. Elle quantifie combien c'est étonnant de découvrir qu'une règle possède un petit nombre de contre-exemples quand l'ensemble de données est important [GR17].

1.2.2 Modélisation mathématique de l'approche de base

Notons A et B les sous-ensembles respectifs de E d'individus qui vérifient respectivement les variables booléennes a et b (Figure 11). $\overline{A}$ et $\overline{B}$ sont les ensembles complémentaires de A et B respectivement dont les cardinaux sont : $card(E) = n$, $card(A) = n_a$, $card(B) = n_b$, $card(\overline{A}) = n_{\overline{a}} = (n - n_a)$, $card(\overline{B}) = n_{\overline{b}} = (n - n_a)$.

Pour une règle quelconque $a \rightarrow b$, observée dans E, l'ASI prend plutôt en considération la non-satisfaction de l'implication $a \Rightarrow b$ qui, apparaît dès lors que a étant vrai, b est faux. Elle représente le nombre de contre-exemple $n_{a \wedge \overline{b}}$ à cette règle observée dans l'intersection $A \cap \overline{B}$. L'ASI consiste à comparer le nombre de contre-exemples $n_{a \wedge \overline{b}}$ avec le nombre de contre exemples qui apparaîtraient lors d'un choix aléatoire et indépendant de deux parties de mêmes cardinaux respectifs que A et B (Figure 11) [Gra79]. Pour formaliser l'hypothèse que a et b sont indépendants, les auteurs ont considéré, comme I.C. Lerman dans [Ler], deux parties quelconques X et Y de E, choisies aléatoirement et indépendamment (absence de lien a priori entre ces deux parties) et de mêmes cardinaux respectifs que A et B. Soit $\overline{Y}$ et $\overline{B}$ les complémentaires respectifs de Y et de B dans E de même cardinal [GKG15][GCG15]. Soit α un réel quelconque de l'intervalle [0,1].

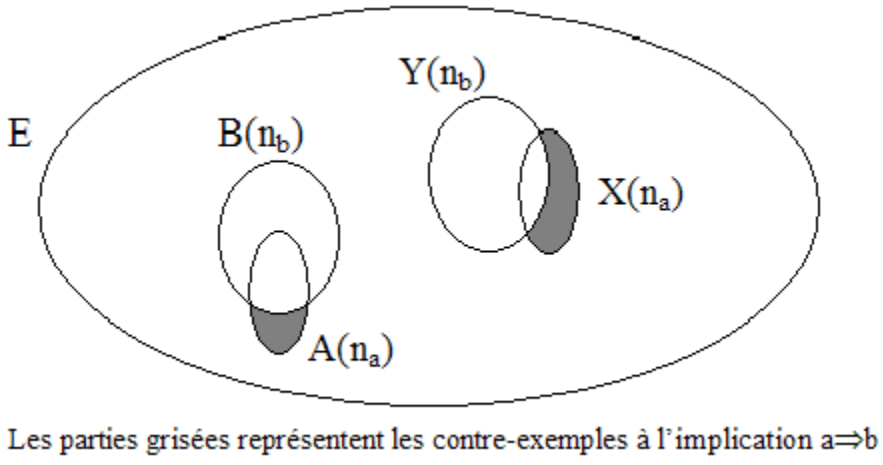

FIGURE 1.1 – Représentation par les diagrammes d'Euler

Définition 1 la quasi-règle $a \Rightarrow b$ est admissible au niveau de confiance $1 - \alpha$ si et seulement si :

$$Pr[Card(X \cap \overline{Y}) \leq card(A \cap \overline{B})] \leq \alpha$$

Définition 2 : On appelle intensité d'implication de la quasi-règle $a \Rightarrow b$, le nombre

$$\phi(a,b) = 1 - Pr[Card(X \cap \overline{Y}) \leq card(A \cap \overline{B})] \text{ si } n_b \neq n \text{ et } \phi(a,b) = 0 \text{ si } n_b = n$$

Cet indice permet de mesurer l'étonnement (une surprise) dû aux faits que le nombre de contre-exemples à la règle $a \Rightarrow b$ est petit par rapport aux grands nombres d'instances, alors que a et b sont supposés indépendants.

La détermination de l'intensité d'implication dépend du modèle retenu pour définir la modélisation de la loi de tirage de X et de Y [GR13]. Si le tirage des transactions est fait une à une, $Card(X \cap \overline{Y})$ suit une loi de Poisson, avec laquelle on obtient l'indice de base pour des variables binaires. La modélisation retenue avec l'ASI est la modélisation binomiale [Gra79] quant X et Y sont tirés avec remise. Lorsque le nombre total de transaction devient très grand les trois modélisations convergent vers le même modèle Gaussien (à titre de généralisation) [Ler]. L'indice de base pour les variables binaires (poison) est le suivant :

$$\phi(a,b) = \frac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{+\infty} \exp(\frac{-t^2}{2}) \mathrm{d}t \tag{1.1}$$

L'intensité d'implication est une valeur probabiliste, contrairement aux autres indices implicatifs les plus utilisées, qui consiste simplement à retenir ou non la quasi-implication entre les variables binaires a et b.

L'ASI distingue les cas triviaux avec pertinence contrairement aux autres méthodes du fait que si B est très grand ou égal à n l'intensité d'implication devient très faible, voire nulle [GA96]. Donc Les règles triviales qui sont potentiellement évidentes et connues de l'expert ne sont pas retenus.

1.2.3 Extension de l'ASI

Différentes adaptations de l'intensité d'implication ont été proposées en concevant à chaque étape les notions mathématiques en réponse aux problèmes posées : entre variables non binaires, c'est-à-dire modales ou fréquentielles [Lag98] [SGK00], entre variables-sur-intervalles, variables-intervalles, variables floues, [Gra00],[LR98] et entre des classes de variables de nature quelconque. La problématique classique a été également élargie à la découverte de règles de type « si a alors presque b » à la recherche de règles généralisées de type R $\Rightarrow$ R' où les prémisses R et les conclusions R' peuvent être elles-mêmes des règles. Une première formalisation basée sur la notion de « hiérarchie orientée » a été proposée par [GKB01]. Une nouvelle formalisation du modèle a mis plus nettement en évidence les structures en jeu [GKB03]. L'ASI a été appliquée dans plusieurs domaines, ce qui a mis en évidence la nécessité d'autres développements théoriques en écho aux questions posées par les praticiens, tels que les notions de typicalité et de contribution des variables supplémentaires, et le développement d'autres indices (l'indice d'implication inclusif, entropique et d'indice d'implifiance).

Les variables supplémentaires ne contribuent pas au calcul des relations impliquées dans la hiérarchie implicative, mais elle apporte une information supplémentaire pour son interprétation (âge, sexe). Les concepts comme la contribution et la typicalité des individus et des variables dans la formation de classes pour la classification des similarités, ou bien dans la hiérarchie cohésive, aident à interpréter et comprendre la nature des rapports entre individus et variables. Elles offrent de nouvelles opportunités d'interprétation didactique des résultats obtenus. Le concept de typicalité représente la proximité des sujets avec le comportement moyen de la population envers les règles statistiques extraites, en d'autres termes, le comportement de ces sujets est ainsi en harmonie avec le comportement statistique de la population à l'origine de la classe. La contribution quantifie le rôle qu'ont les sujets par rapport aux règles strictes associées [GDRG06]).

1.2.4 Autres indices

Dans cette section nous présentons les limites de la mesure d'intensité d'implication et les principales mesures utilisées par l'ASI pour remédier a ces limites. Nous présentons d'abord l'indice d'inclusion ensuite l'indice entropique et finalement la mesure d'implifiance.

1.2.4.1 Limites de la mesure d'intensité d'implication

L'intensité d'implication présente l'inconvénient d'être peu discriminante quand les cardinaux étudiés sont grands, car ses valeurs peuvent être souvent proches de 1 alors que A n'est pas inclus dans B [GKCG01]. D'où la nécessité d'adapter le concept d'intensité à des situations où les populations en jeu deviennent très importantes. Pour résoudre ce problème, Gras et al ont proposé dans [GKCG01] de moduler les valeurs de l'intensité d'implication par un indice de quasi-implication fondé sur l'entropie de Shannon : l'indice d'inclusion. L'indice formé s'appelle intensité d'implication entropique. Les utilisateurs de cet indice ont apprécié la capacité à accepter plus facilement la grande taille de l'échantillon des sujets considérés. D'où son intérêt pour ce que l'on appelle les « big data ». Ce dernier présente aussi un caractère jugé trop ad-hoc par les familiers de l'ASI [GCG15]. Ceci a motivé les auteurs à créer un nouvel indice appelé implifiance. Tous ces indices prennent en compte la contraposée $\overline{B} \Rightarrow \overline{A}$ qui permet de renforcer l'affirmation de la relation implicative de a sur b. Elle pourrait également contribuer à répondre aux problèmes de l'approche support-confiance puisque si on a un support très petit avec une confiance très élevé c-a-d si A et B sont petits relativement à E leurs complémentaires seront grands et réciproquement [GKCG01].

1.2.4.2 Indice d'inclusion

L'indice d'inclusion, est un indice descriptif fondé sur l'entropie. Gras a développé cet indice spécialement pour l'associer à l'intensité d'implication afin de pallier ces insuffisances. Cette nouvelle mesure permet de mieux modéliser la qualité de l'inclusion entre le sous-ensemble d'individus

vérifiant A et celui vérifiant B et de prendre en compte à la fois les contre-exemples de $A \Rightarrow B$ et de sa contraposée, il s'agit de moduler la valeur de l'étonnement en fonction du :

— Déséquilibre entre $n_{a \wedge b}$ et $n_{a \wedge \overline{b}}$ associé à $A \Rightarrow B$

— Déséquilibre entre $n_{a \wedge \overline{b}}$ et $n_{\overline{a} \wedge \overline{b}}$ associé à $\overline{B} \Rightarrow \overline{A}$

Une mesure bien connue pour évaluer les déséquilibres de façon non linéaire est l'entropie de Shannon [Sha01]. L'entropie conditionnelle $H_{B/A}$ relative aux cas (A et B) et (A et $\overline{B}$) lorsque A est vérifié est définie dans : [GKCG01].

$$H_{B/A} = -\frac{n_{A \wedge B}}{n_A} \log \frac{n_{A \wedge B}}{n_A} - \frac{n_{A \wedge \overline{B}}}{n_A} \log \frac{n_{A \wedge \overline{B}}}{n_A}$$

De même, l'entropie conditionnelle $H_{\overline{A}/\overline{B}}$ relative aux cas ($\overline{A}$ et $\overline{B}$) et (A et $\overline{B}$) lorsque $\overline{B}$ est vérifié est définie par :

$$H_{\overline{A}/\overline{B}} = -\frac{n_{A \wedge \overline{B}}}{n_{\overline{B}}} \log \frac{n_{A \wedge \overline{B}}}{n_{\overline{B}}} - \frac{n_{\overline{A} \wedge \overline{B}}}{n_{\overline{B}}} \log \frac{n_{\overline{A} \wedge \overline{B}}}{n_{\overline{B}}}$$

D'une façon générale, ces entropies devraient être simultanément petites si l'on souhaite disposer d'un bon critère d'inclusion de I_A dans I_B. Cependant, l'impact des déséquilibres doit être ajusté en fonction des différentes situations cardinales. L'étude des courbes de l'entropie conditionnelles a conduit à de nouvelles mesures d'inclusions h_1 et h_2 associée respectivement aux règles :$A \Rightarrow B$ et $\overline{B} \Rightarrow \overline{A}$.

L'indice d'inclusion ainsi obtenu est : $\iota_{AB} = (1 - h_1)(1 - h_2)^{1/2}$. Il rassemble les informations obtenues des deux quasi règles $A \Rightarrow B$ et $\overline{B} \Rightarrow \overline{A}$ [GKCG01]

1.2.4.3 L'intensité Entropique

L'association de l'intensité d'implication et de l'indice d'inclusion crée un indice de quasi-implication, nommée intensité d'implication entropique, qui est de nature statistique (grâce à l'intensité d'implication) tout en restant discriminant quand les cardinaux étudiés sont grands (grâce à l'indice d'inclusion). L'association des deux mesures est réalisée par la moyenne géométrique [GKCG01].

Définition. L'intensité entropique de la règle $A \Rightarrow B$ est définie par $\Psi_{AB} = (\phi_{AB}.\iota_{AB})^{1/2}$ où ϕ_{AB} est l'intensité d'implication et ι_{AB} l'indice d'inclusion.

Cette version "entropique" permet de mieux modéliser l'inclusion et de prendre en compte les contre-exemples à $A \Rightarrow B$ et les contre-exemples à sa contraposée. De ce fait, elle est peu sensible aux bruits et ne varie pas linéairement avec les cardinaux des sous-ensembles en jeu.

1.2.4.4 Implifiance

Vu que le critère de confiance ne varie pas pour toute dilatation de E et des exemples de A et de B, et que la confiance devient maximale lorsque le nombre de sujets vérifiant b tend vers n ou égal

à n, ce dernier a été refusé par certaines auteurs comme seul critère de révélation d'une relation implicative. L'intensité d'implication devient triviale dans ce cas, mais présente l'inconvénient d'être peut discriminante quand les cardinaux étudiés sont grands.

La contraposée $\overline{B} \Rightarrow \overline{A}$ permet de renforcer l'affirmation de la relation implicative de A sur B. Elle contribue à résoudre le problème dont souffre l'approche support-confiance (le cas des règles ayant un support trés petit avec une confiance élevé). Car si A et B sont petits relativement à E leurs complémentaires seront grands et réciproquement [GKCG01].

Les auteurs ont combiné l'intensité d'implication avec la confiance et la contraposée de l'implication pour une nouvelle modélisation de l'implication entre deux variables binaires. L'indice ainsi formé est appelé « implifiance ».

La valeur de confiance de la règle $A \Rightarrow B$ et $\overline{B} \Rightarrow \overline{A}$ sont respectivement $C1$ et $C2$:

$$C_1(a,b) = Fr[Y|X] = ([cardX \cap Y])/(cardX) = \frac{n_{a\wedge b}/n}{n_a/n} = \frac{n_{a\wedge b}}{n_a}$$

$$C_2(\overline{b},\overline{a}) = Fr[\overline{X}|\overline{Y}]) = ([card\overline{X} \cap \overline{Y}])/(card\overline{Y}) = \frac{n_{\overline{a}\wedge\overline{b}}/n}{n_{\overline{b}}/n} = \frac{n_{\overline{a}\wedge\overline{b}}}{n_{\overline{b}}}$$

Les deux confiances sont différentes en général. Donc C_1 et C_2 ont été utilisées comme indicateurs des inclusions partielles : A inclus dans B et B inclus dans A. Par suite, ils ne permettront pas de déceler des relations « surprenantes » eu égard à l'échantillon observé de taille n. C'est pour cette raison que l'intensité d'implication classique est associée à la confiance.

On appelle implifiance la mesure de l'implication statistique qui prend en compte l'implication directe et sa contraposée, ainsi que la confiance en chacune de ces deux formes inclusives. Sa valeur est :

$$\Phi(a,b) = \phi(a,b).[C_1(a,b).C_2(\overline{b},\overline{a})]^{1/4} \tag{1.2}$$

Par rapport à l'intensité entropique, l'implifiance est moins complexe. Elle est aussi bien définie dans les cas où les variables sont de nature quelconque (numériques, modales, floues, variables intervalles, etc.)

1.3 (R)CHIC

CHIC est un outil informatique, permettant d'utiliser la plupart des méthodes définies dans le cadre de l'ASI, il a été écrit en C++, à partir d'une ancienne version écrite en Pascal mais avec beaucoup d'autres développements et une interface plus conviviale [Cou00]. La version actuelle de ce logiciel appelé RCHIC écrite par Raphaël Couturier a été portée en R [Cou22], fonctionnel sous Windows et Linux et MacOS conçu à partir de la version en C++. Cette dernière version subie régulièrement des mises à jour, ce qui le met au même niveau avec les déférents développements théoriques de l'ASI. Ce logiciel a pour objectif de découvrir les implications les plus pertinentes

entre les variables d'un ensemble de données. Il permet de construire deux types de hiérarchie et un graphe.

1.3.1 Données/ Variables

(R)CHIC traite les données disposées sous forme d'un tableau numérique enregistré dans un fichier CSV. Les individus sont rangés dans la première colonne du tableau. Les variables sont disposées sur la première ligne. Les valeurs des individus sont représentées dans un tableau à deux dimensions tel que les valeurs pour chaque variable d'un individu sont rangées dans une ligne du tableau (le premier élément étant le nom de l'individu). Les valeurs d'une variable pour tous les individus sont disposées dans les colonnes du tableau (le premier élément étant le nom de la variable). Voici un extrait d'un fichier contenant des données traitées par le logiciel (R)CHIC.

	Affectueux	Agile	Agressif	Angoissant	Attirant	Beau
Aigle	0	1	0	1	1	1
Ane	0	0	0	0	0	1
Autruche	0	0	1	1	0	0
Baleine	0	0	0	1	1	1
Bouc	0	1	1	0	1	0
Canard	0	0	1	0	0	1
Chamois	0	1	0	0	0	1
Chat	1	1	0	0	1	1
Chien	1	0	0	0	0	0
Cigale	0	0	0	0	0	1
Corbeau	0	0	1	1	0	1
Couleuvre	0	0	0	1	0	0
Crocodile	0	0	1	1	0	0

FIGURE 1.2 – Extrait d'un fichier contenant des données traitées par le logiciel (R)CHIC

CHIC offre la possibilité de traiter différents types de variable, ces variables peuvent être principales ou supplémentaires selon leurs interventions dans les calculs. Principales si elles interviennent directement dans tous les calculs, sinon elles sont supplémentaires. CHIC traite les variables : binaire, modale, fréquentielle, quantitative ou intervalle. Le cas des variables binaire est le cas le plus simple. Les variables fréquentielles quant à elles prennent leurs valeurs entre 0 et 1. Ce type de variable permet de modéliser les variables modales pour lesquelles il existe un nombre fixe de valeurs comprises entre 0 et 1 qui correspondent aux différentes modalités. Les valeurs des variables quantitatives sont transformées en variables fréquentiels dans l'intervalle [0-1] en divisant toutes les valeurs par la valeur maximum obtenue par la variable. Les variables-intervalles découpe les valeurs de la variable en différents intervalles par un algorithme approprié, de type « nuées dynamiques » [Did71], qui, à partir d'un nombre d'intervalles choisi par l'utilisateur, constitue des intervalles en minimisant l'inertie de chaque intervalle et en maximisant l'inertie interclasse de l'ensemble des

intervalles. Ensuite, un intervalle est représenté par une variable binaire et un individu a la valeur 1 s'il appartient à cet intervalle et 0 sinon. Les contributions de cette thèse traitent les variables d'intervalles supplémentaires. (R)CHIC traite ces données avec l'algorithme apriori pour former toutes les implications et calcule pour chaque implication : le nombre d'occurrence, le support, la confiance, l'indice d'implication, l'indice entropique ... etc. toutes ces informations sont enregistrées sous forme d'un tableau dans un fichier appelé transaction.out. Voici un extrait de ce fichier.

hyp -> con	occurrence	occurrence	supporti	confidence	classical index	entropic index	classical simi
Vif -> Laid	24.0000000000	21.000000000	58.5365853	45.8333333333	28.797902807798	18.22523905725300	35.61771512031
Laid -> Vif	21.0000000000	24.000000000	51.2195121	52.3809523809	25.992417099598	17.71341488256969	35.61771512031
Vif -> Mecha	24.0000000000	21.000000000	58.5365853	50.0000000000	39.060149003760	26.12109595719913	46.67355120182
Mechant -> \	21.0000000000	24.000000000	51.2195121	57.1428571428	37.409122748026	26.93880200106020	46.67355120182
Vif -> Crainti	24.0000000000	20.000000000	58.5365853	50.0000000000	45.746898209761	31.12504281893171	53.40839624404
Craintif -> Vi	20.0000000000	24.000000000	48.7804878	60.0000000000	44.823575957166	33.65059212178595	53.40839624404
Vif -> Beau	24.0000000000	20.000000000	58.5365853	50.0000000000	45.746898209761	31.12504281893171	53.40839624404
Beau -> Vif	20.0000000000	24.000000000	48.7804878	60.0000000000	44.823575957166	33.65059212178595	53.40839624404
Vif -> Sournc	24.0000000000	20.000000000	58.5365853	58.3333333333	68.268933347263	50.75696707982637	74.85910654067
Sournois -> \	20.0000000000	24.000000000	48.7804878	70.0000000000	72.105253363508	59.15307940535562	74.85910654067
Vif -> Gentil	24.0000000000	19.000000000	58.5365853	33.3333333333	15.586446936562	8.558525599976480	17.46034175157
Gentil -> Vif	19.0000000000	24.000000000	46.3414634	42.1052631578	10.332223783077	6.415100768088446	17.46034175157
Vif -> Discret	24.0000000000	19.000000000	58.5365853	50.0000000000	52.342726257786	36.14041788247207	60.38349866867
Discret -> Vif	19.0000000000	24.000000000	46.3414634	63.1578947368	52.989073719006	41.36957018739217	60.38349866867
Vif -> Agress	24.0000000000	19.000000000	58.5365853	41.6666666666	31.247959963371	19.49555869935612	36.82766556739

FIGURE 1.3 – Extrait du fichier transaction.out

1.3.2 Graphe implicatif

Le graphe implicatif représente les implications précédemment calculées dans «transactions.out » dont l'intensité d'implication est supérieure à un certain seuil. Une implication est représentée par une flèche qui relie deux variables. (R)CHIC permet de sélectionner quatre seuils identifiés par quatre couleurs différentes. L'utilisateur choisit le seuil désiré. La figure suivante représente un graphe d'implication avec le menu pour le choix des seuils :

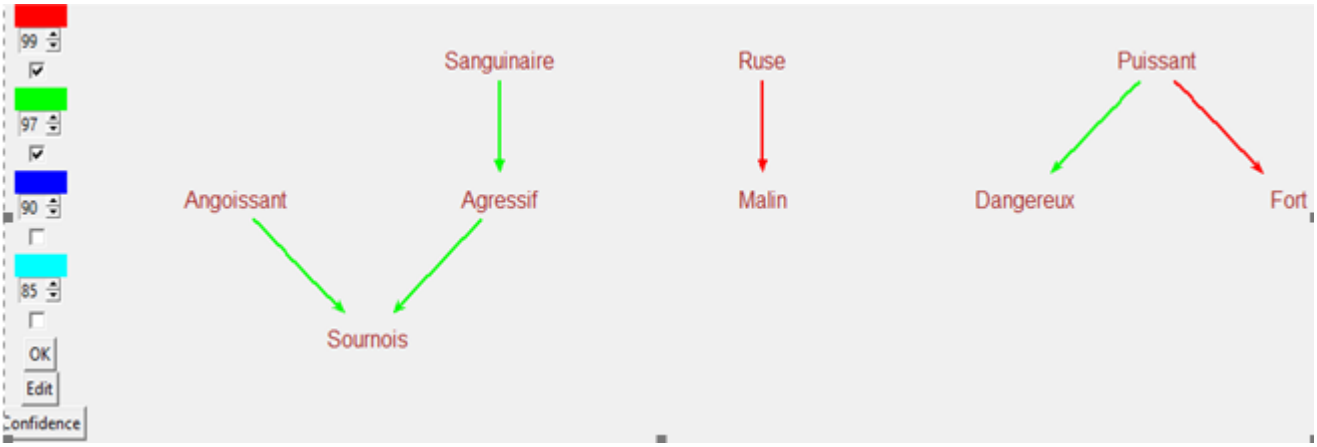

FIGURE 1.4 – Exemple d'un graphe implicatif

Par défaut toutes les variables impliquées dans le graphe sont représentées dans la zone de travail. Puis au cours de l'interprétation, l'utilisateur peut se rendre compte que seules certaines variables lui semblent utiles. Dans ce cas, il supprime temporairement les variables désirées grâce à une boite de dialogue prévu à cet effet. (R)CHIC met à jour à nouveau le graphe des implications. À tout moment il est possible d'ajouter ou de supprimer des variables dans l'analyse que l'on effectue. La figure suivante représente la boite de dialogue.

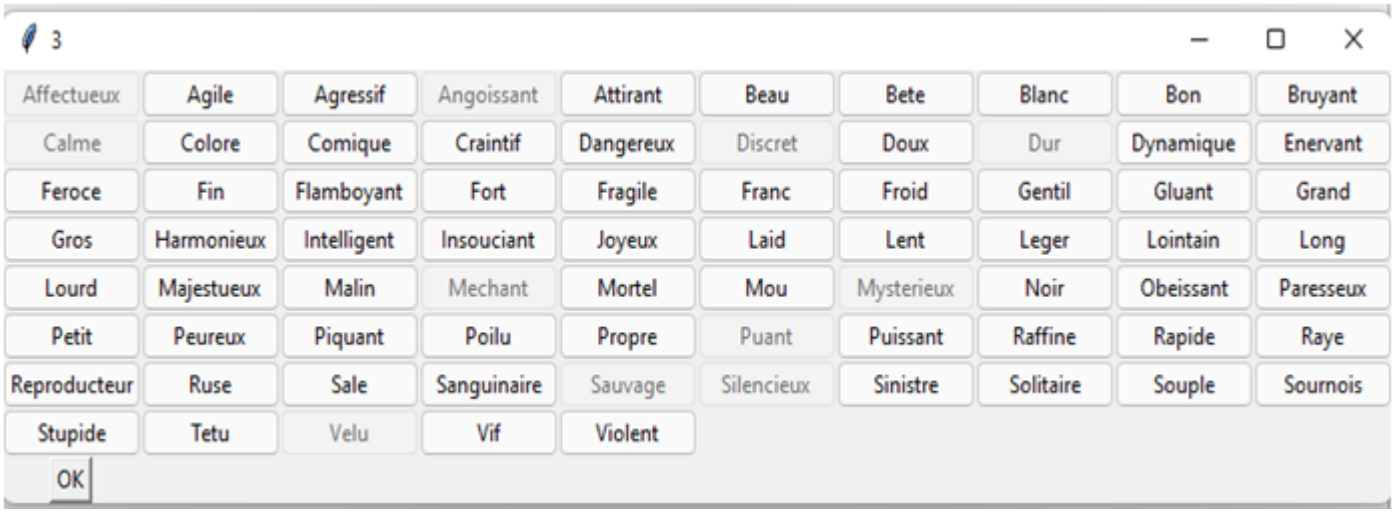

FIGURE 1.5 – Exemple d'une boite de dialogue pour l'ajout ou la suppression d'une variable.

RCHIC offre aux utilisateurs plusieurs modes pour le calcul, les modes existants sont : l'implication standard ou intensité d'implication (index classique), l'indice entropique, l'intensité d'implication + confiance (Voir Chapitre 3) et l'implifiance.

Par défaut les fermetures transitives ne sont pas affichées sur le graphe implicatif afin de minimiser le nombre de croisements. CHIC offre la possibilité de les faire apparaitre, et cela par un simple clic sur la souris dans la boite à outils. Il est aussi possible de sauvegarder l'état d'un

graphe ou même plusieurs états sur le même graphe et ainsi mettre en évidence différentes parties du graphe. CHIC offre aussi la possibilité d'exporter un graphe sous Word ou Excel.

1.3.3 Arbre cohésif

Afin de construire l'arbre cohésitif, nous avons toujours besoin de l'ensemble de toutes les implications (transaction.out) et des valeurs d'intensité d'implications. Les implications sont reliées entre elles selon leurs cohésions pour former des classes (l'implication est la forme la plus simple d'une classe). Pour former la hiérarchie, à chaque niveau de classification, (R)CHIC choisit la classe qui possède la plus grande cohésion (en termes d'intensité d'implication), et à chaque étape (R)CHIC agrège une classe existante avec soit une variable qui n'as pas été agrégée soit avec une autre classe conduisant à la cohésion la plus forte a cette étape, ce qui permet de former une nouvelle classe [CA09].

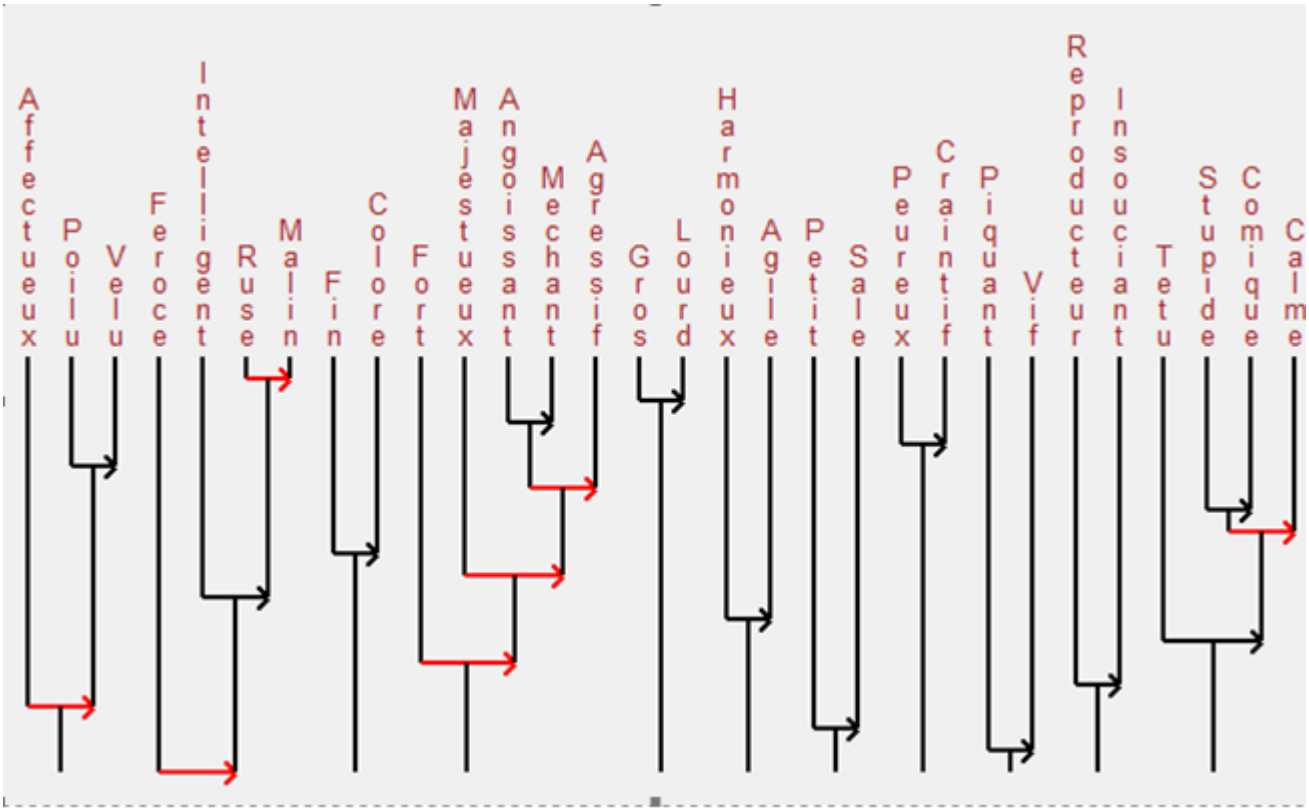

FIGURE 1.6 – Exemple d'un arbre cohésitif

1.3.4 Arbre des similarités

L'arbre des similarités est la hiérarchie la plus connue, elle utilise la liste des implications et l'indice de similarité classique défini dans [Ler], pour agréger les classes. Dans le cas où l'utilisateur traite un grand nombre d'individus, (R)CHIC offre la possibilité d'utiliser la similarité entropique pour tracer l'arbre. L'arbre calcule pour chaque couple de variables la similarité entre celles-ci.

Ensuite, il agrège des classes constituées elles-mêmes d'autres classes. Les niveaux identifiés par un trait rouge sont les niveaux les plus significatifs par rapport aux autres niveaux [CG05]

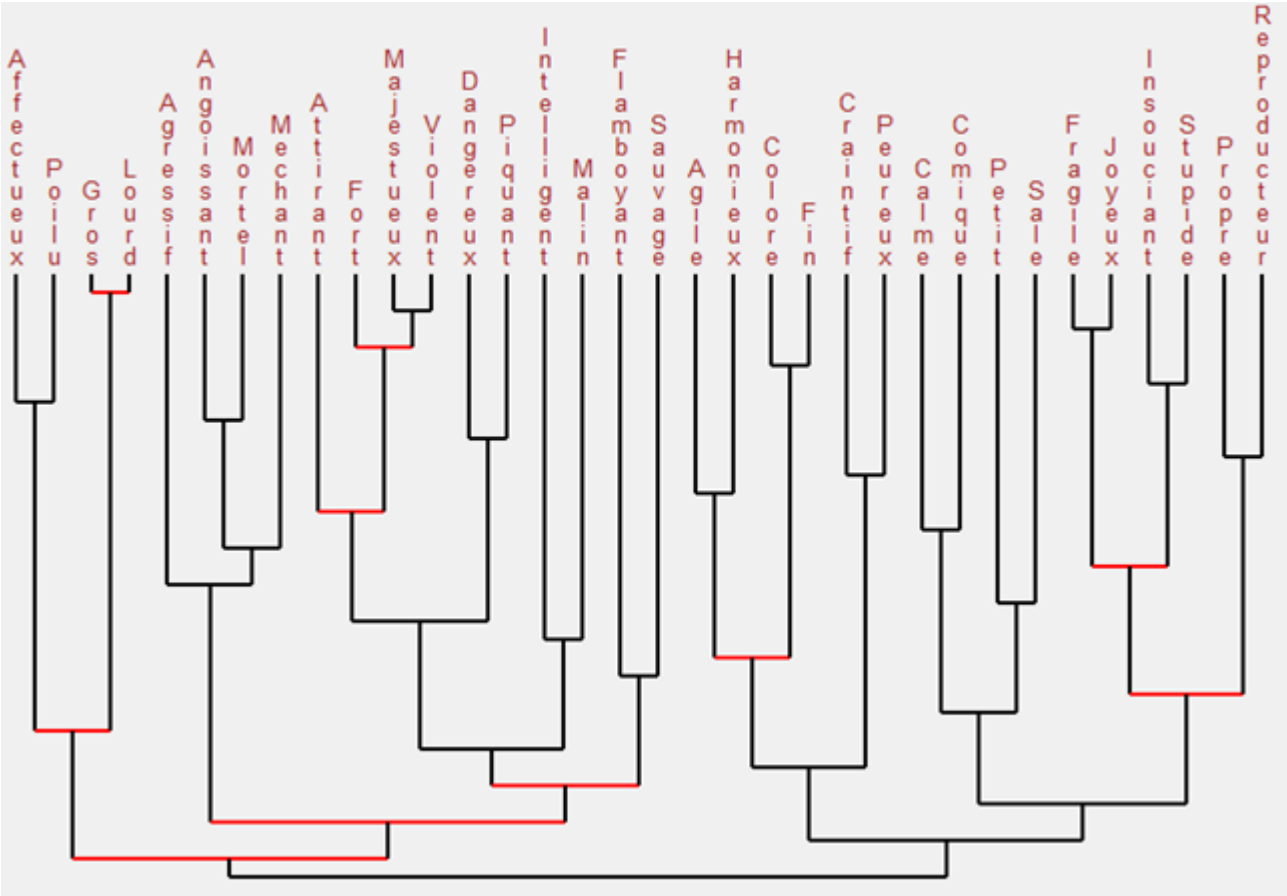

FIGURE 1.7 – Exemple d'un arbre de similarité

La théorie classique conduit à la fin du processus à une seule classe qui rassemble toutes les autres, contrairement à la théorie entropique qui conduit à plusieurs classes distinctes, dont le nombre de classes dépend de la similarité des données [CA09]

1.4 Domaines d'application de l'ASI

CHIC et l'ASI ont été utilisés pour un large spectre de domaines de recherche, en psychologie, éducation, marketing, médecine ...etc. Nous présentons quelques approches qui utilisent l'ASI à travers le logiciel (R)CHIC,

Dans [Ram08], les auteurs introduisent l'utilisation des concepts de l'Analyse Statistique Implicative pour analyser les données issues de puces à ADN. Les puces à ADN permettent l'analyse simultanée de plusieurs milliers de gènes. Le jeu de données biologiques est constitué de 7129 gènes et de 38 patients. L'analyse des puces à ADN peut conduire à la prédiction de deux types de leucémie dont souffre le patient. Les classes de leucémie sont : ALL (Acute Lymphoblasic Leukemia) présente 27 individus dans le jeu de données et AML (Acute Myeloid Leukemia) présente

11 individus. L'objectif ne se résume plus uniquement à guérir des patients, mais à diagnostiquer des risques futurs de la maladie. Le but était d'identifier les deux groupes de patients existants et cela à travers l'analyse implicatif du logiciel CHIC et d'extraire les gènes qui différencient le mieux ces groupes. Les données des puces à ADN regroupent plusieurs milliers de gènes mais ne portent que sur un nombre limité d'expériences en raison du coût élevé du procédé. Cela pose un véritable problème pour l'extraction de règles d'association pour définir des associations entre gènes. Les auteurs proposent dans cette approche une méthodologie permettant de réduire le nombre de gènes étudiés en utilisant un critère simple qui consiste à ne retenir que les gènes qui s'expriment au moins k fois sur les n expériences. Avec l'ASI les auteurs peuvent annoncer qu'une implication est valide avec une mesure de qualité donnée, mais le plus important pour les biologistes est le fait qu'avec l'ASI on peut déterminer les entités concernées réellement par ce résultat en utilisant le concept de contribution qui consiste à chercher l'individu le plus contributif dans chacune des classes.

L'approche présentée dans [OGC05] montre une utilisation didactique des variables supplémentaires et des élèves fictifs à travers le logiciel CHIC. L'objectif initial des enseignants du département de mathématique de l'université Jaume I de Castelló (UJI) en Espagne est de constater et d'analyser les connaissances en mathématiques de leurs étudiants par rapport à ce qu'ils supposent acquis à la fin de l'enseignement secondaire. Les auteurs ont collecté des données sur leurs étudiants auxquelles ils ont rajouté la variable TOTAL, qui attribue à chaque individu le nombre de réponses correctes comme indicateur d'un niveau de connaissances en mathématique. Les résumés statistiques de cette variable ont montré des différences significatives entre les niveaux de la réussite globale des étudiants en fonction des types d'études. Le calcul de la contribution des variables supplémentaires, montre que l'unique variable supplémentaire qui contribue à la formation de la classe la plus significative de l'analyse des similarités, est la variable "avoir suivi récemment un cours de préparation de Mathématiques". Elle en est également typique. Les tests ont été appliqués par la suite sur des données fictives où l'analyse de similarité et de cohésion, ont permis de confirmer la caractérisation des classes significatives, et que le changement d'occurrences produit par l'introduction de ces élèves dans l'échantillon, ne modifie pas très sensiblement les résultats globaux des analyses abordées.

Dans [Cou15] l'auteur propose une nouvelle méthodologie pour mettre en place un système de recommandation basé sur l'ASI. Contrairement aux autres approches ou l'utilisateur est obligé de se dévoiler avant d'avoir des recommandations, cette approche offre la possibilité d'établir des recommandations sans avoir un profil utilisateur. Ceci est assurée du fait que l'auteur ne prend pas l'avis de l'utilisateur qu'on veut recommander, mais plutôt utilise des données recueillies à partir de critiques d'un cite de la presse. Ces critiques ont été recueillis et placé dans une base de données. La base comporte 16657 avis sur un total 1248 films critiqués par 49 revues de presse. Les avis des utilisateurs dans le cite sont représentés par des étoiles (1 à 4). À partir de ces données, CHIC

calcule les implications et garde que celles supérieurs à un certain seuil. Les auteurs jugent que l'indice entropique est plus approprié que les autres indices. Une implication s'interprète comme ceci : si la presse a apprécié un film donc elle a également apprécié une liste d'autres films. Et les recommandations se font selon qu'un utilisateur choisit de regarder ce film ou pas.

Dans [DGB+08], les auteurs ont répondu au besoin de l'Association pour l'Emploi des Cadres (APEC) d'élaborer des référentiels comportementaux destinés à faciliter la réinsertion à travers un système d'aide à la décision destiné aux consultants en charge de l'accompagnement au repositionnement professionnel. Les bases psychologiques des outils déjà existants ont toutes été validées, mais très peu d'entre eux ont fait l'objet d'une analyse statistique approfondie. C'est le cas de l'outil d'évaluation PerformanSe Echo qui donne le profil comportemental d'une personne selon 10 dimensions bipolaires. Chacune des 10 dimensions du modèle est représentée par trois variables suivant que le sujet vérifie peu (-), moyennement (0), beaucoup (+) cette dimension. Comme exemple de comportement nous citons Affirmation/Remise en cause (AFF/RMC), l'affirmation qui exprime la confiance en soi. Echo a été validé sur une population de 4538 sujets en 2004. Dans cette approche les auteurs s'intéressent à la construction d'un ensemble d'indicateurs psychologiques basés sur Echo sur une population de 613, âgés de 45 ans et plus, et en recherche d'emploi. Cette étude est réalisée à l'aide du logiciel CHIC. La méthodologie proposée se décompose en deux temps : tout d'abord une analyse globale de l'échantillon qui mettra en évidence des combinaisons de dimensions caractéristiques, puis une analyse plus approfondie de certaines de ces combinaisons pour qualifier des indicateurs significatifs pour un consultant. Et cela en se servant de l'arbre des similarités, l'arbre cohésitif et du graphe implicatif. Les auteurs ont identifié trois groupes principaux sur la dimension affirmation (AFF). Il apparaît que les indicateurs trouvés étaient tout à fait pertinents au regard du comportement global de chacun de ces trois groupes. En effet, chacun de ces groupes à un comportement particulier compte tenu de sa réinsertion dans le monde du travail. Le groupe en Affirmation faible se caractérise principalement par une période de réinsertion plus longue et par un sentiment de défaite par rapport à sa situation de chômage, alors que le groupe en Affirmation forte a un taux de réussite plus élevé en réinsertion et montre un comportement plus positif vis-à-vis de sa situation. Le groupe en affirmation moyenne est moins clairement défini que les deux autres et le comportement de ses sujets est moins uniforme. Certaines d'entre eux suivent la tendance du groupe AFF-, d'autres celle de l'AFF+. Cette étude a conduit à des découvertes intéressantes selon l'expert psychologue. D'abord l'approche a montré que la méthode Echo précédemment utilisée pourrait donner des résultats imprécis, et dans certains cas, erronés. Ainsi, cette étude basée sur l'ASI a prouvé l'intérêt d'utiliser des outils statistiques et des méthodes d'analyse de données plus avancés dans le domaine de la psychologie. En effet, l'étude montre que CHIC peut être utilisé comme un outil d'aide à la décision, combiné à l'outil d'évaluation psychologique Echo.

Dans [DR17] les auteurs ont présenté quelques résultats issus d'une recherche portant sur les difficultés et facilités d'apprentissage de la statistique. Ils évaluent les connaissances et compétences

des étudiants en utilisant quelques notions élémentaires de statistique. Telles que : moyenne, écart-type et intervalle de confiance afin de repérer leurs niveaux de conceptualisation et mieux comprendre les obstacles auxquels ils se confrontent. Les données construites au moyen d'une enquête par questionnaire sont traitées dans un premier temps selon une approche statistique classique. Puis en utilisant le cadre théorique de l'analyse statistique implicative à travers l'usage du logiciel C.H.I.C. La mise en œuvre de l'ASI a permis de dégager de résultats complémentaires qui enrichissent ceux issus de l'analyse descriptive initiale.

Dans [BBT10] les auteurs se sont intéressés à l'évolution des modalités de formation et de construction des professions des enseignants débutant. Les auteurs ont utilisé des questionnaires composés de 28 items convergeant avec le référentiel métier publié en 2007. 900 stagiaires de Formation des Maîtres ont répondu à ce questionnaire. Les enseignants censés répondre tiennent compte de la construction de chaque compétence, est-elle importante ? Faisable ? ou en cours de leur pratique. L'utilisation de l'analyse statistique implicative a permis de mettre en évidence une conception unifiée de la profession idéalisée chez les enseignants des premiers et seconds degrés mais une structuration plus grande chez les professeurs de lycée et collège de leur pratique par des échanges au sein de collectifs professionnels pour traiter les problèmes de classe.

Dans [KBC17] Les auteurs ont tenté d'appliquer l'ASI aux données PISA (Programme International pour le Suivi des Acquis des élèves). Le travail sur ces données a commencé au milieu de l'année 1990 en réponse aux demandes des pays membres de l'organisation (OCDE) (Organisation de Coopération et de Développement Economique), pour le suivi des données régulières et fiables sur les connaissances et les compétences de leurs élèves et la performance de leurs systèmes éducatifs. L'enquête se fait tous les trois ans. Les données traitées par les auteurs sont les données du domaine littératie mathématique de PISA2012. La méthode consiste à analyser les dépendances entre les questions de l'enquête et de comparer le savoir des élèves selon les pays participant à l'enquête. Les auteurs ont utilisé les graphes (graphe implicatif et arbre cohésif) réalisés à partir du logiciel RCHIC pour montrer les relations entre les questions. L'utilisation de RCHIC a permis aux auteurs de traiter une masse importante de données de plusieurs pays. De plus, la méthode peut être utilisée pour d'autres domaines du même type. Les auteurs ont prouvé que l'utilisation de l'ASI permet de montrer les liens entre les différentes questions de test PISA et de montrer que la réponse à une question donnée implique parfois la réponse à une autre question.

1.5 Classification

De grandes quantités de données ne cessent d'être produites partout dans le monde, vu l'abondance de la capacité de stockage et la rapidité de transmission des réseaux. Cependant, si ces données ne sont pas exploitées leur collecte deviendrait inutile. Le Data Mining est un domaine permettant, à partir d'une très importante quantité de données brutes, d'en extraire de façon automatique ou semi-automatique des informations cachées, pertinentes et inconnues auparavant.

Dans ce but plusieurs techniques ont été développées pour répondre à ce problème. On distingue les techniques de Data Mining supervisé et le Data Mining non supervisé.

L'apprentissage supervisé est une technique d'apprentissage automatique connu sous le terme anglais de machine learning, qui permet à une machine d'apprendre à réaliser des tâches à partir d'une base d'apprentissage contenant des exemples déjà traités. L'apprentissage supervisé concerne essentiellement les méthodes de classification de données. Le but des algorithmes de classification est de construire des catégories ou classes afin de pouvoir mettre tous les éléments des ensembles de données dans ces classes ou catégories.

La classification est un problème crucial pour de nombreuses applications, par exemple la reconnaissance des chiffres manuscrits et le filtre SPAM. Pour le problème de reconnaissance des chiffres manuscrits, l'entrée de l'algorithme de classification est une image numérisée d'un chiffre écrit. Après le traitement (la classification) chaque caractère doit être classé comme l'un des chiffres 0-9 (dix classes en tout). Dans un filtre SPAM (courrier indésirable), chaque message traité doit être classé comme SPAM ou HAM (courrier que l'on désire recevoir). Dans ce dernier, les entrées sont des caractéristiques des messages (telles que : la fréquence de certains mots clés, les majuscules, etc.). En médecine, la confirmation de la présence ou de l'absence d'une maladie, nécessite des informations sur les patients telles que les indicateurs sanguins. Le problème de reconnaissance des chiffres manuscrits et le filtre SPAM sont des exemples de classification binaire. En général la classification est une tâche qui assigne chaque objet à une catégorie prédéfinie. En entrée, nous avons un ensemble de données appelé l'ensemble d'apprentissage, ce dernier comporte des exemples (les classes sont connues à l'avance). L'objectif de l'ensemble d'apprentissage est de construire un modèle de classe de manière à ce que le modèle puisse être utilisé pour classifier de nouvelles données [HPK11]. Le modèle peut être testé avec de nouvelles observations (nouvelles images, nouveaux messages, nouveaux patients). Il permet d'attribuer à chaque observation la classe la plus plausible (c'est-à-dire la lettre ou le chiffre le plus probable, qu'il s'agisse de SPAM ou non, que le patient souffre de la maladie ou non) et sa précision peut être testée par rapport à un ensemble de tests [HPK11].

Parmi les nombreuses méthodes de classification existantes, on peut citer les algorithmes d'arbre de décision, les machines à vecteurs de support, les algorithmes bayésiens, les algorithmes à base de règles, les réseaux de neurones, les méthodes basées sur la distance, les algorithmes génétiques et la classification associative [MMH97].

1.5.1 Présentation des Algorithmes de classification

Nous présentons dans cette section les algorithmes d'apprentissage supervisés les plus populaires, utilisés pour évaluer les deux classificateurs que nous avons proposée dans le chapitre 4 et 5. Nous donnons une brève description de ces algorithmes.

Naive bayes Naïve bayés est un classificateur probabiliste simple basée sur le théorème de

Bayes avec une hypothèse d'indépendance forte (naïve). La prise de décision avec ce classificateur consiste à choisir la meilleure décision possible en se basant sur la loi de bayes et sur le coût associé à chaque décision. Il suppose que l'existence d'une caractéristique pour une classe, est indépendante de l'existence d'autres caractéristiques. Même si ces caractéristiques sont liées dans la réalité. C'est aujourd'hui un algorithme largement utilisée par les outils de Machine learning (capacité donnée aux ordinateurs d'apprendre par eux-mêmes) du fait de ses calculs de probabilités peu coûteux qui lui confèrent une grande agilité [SC95], [Did71]. Dans certains domaines, les performances du classificateur se sont révélées comparables à celles des réseaux de neurones et des arbres de décision.

Arbre de décision CART et J48

Les arbres de décision sont des algorithmes supervisés qui partitionnent récursivement les données en fonction de leurs attributs, jusqu'à ce qu'une certaine condition d'arrêt soit atteinte [FGH+12]. Chaque noeud dans l'arbre (qui n'est pas une feuille) dénote un test sur un attribut, une branche représente le résultat de test, et une feuille une classe prédite. Au niveau de chaque nœud, l'algorithme sélectionne le meilleur attribut pour partitionner les données en classes individuelles. Et cela en faisant appel à des méthodes de sélection d'attributs. Les mesures de sélection sont des heuristiques ou des règles de partitionnement, qui permettent de choisir des critères de partitionnement, afin de former des classes individuelles pures. Les mesures les plus connues sont : Le gain d'information [TW14] : basé sur l'entropie de Shannon, le gain ratio et le Gini index [TW14]. La caractéristique la plus importante des algorithmes d'arbre de décision est leur capacité à décomposer un processus décisionnel complexe en un ensemble de décisions plus simples, fournissant ainsi une solution qui est souvent plus facile à interpréter [Gra79].

CART (Classification And Regression Tree)

CART est un algorithme développé par Breiman, Friedman, Olshen et Stone (1984) [BFOS84]. Cet algorithme est parmi les méthodes de classification les plus connues et les plus utilisées. Il permet de construire un arbre de décision strictement binaire avec exactement deux branches pour chaque nœud de décision. L'ensemble de données est divisé en deux sous-groupes qui sont les plus différents en ce qui concerne le résultat. Le partitionnement est effectué de façon récursive selon la méthode diviser pour mieux régner. CART construit des arbres de décision en utilisant le Gini index comme mesure de sélection d'attribut. Pour chaque nœud de décision, CART fait une recherche exhaustive sur tous les attributs. Pour chaque attribut, le sous ensemble qui donne une valeur minimale de Gini index est sélectionné comme un sous ensemble pour le partitionnement. L'algorithme s'arrête lorsque le nœud est pur (Tous les éléments du nœud appartiennent à la même classe), ou bien tous les attributs ont été utilisés précédemment, ou lorsque la profondeur de l'arbre à atteint la valeur maximale définie par l'utilisateur ou lorsque la taille du nœud est inférieure à la taille minimale définie par l'utilisateur. L'algorithme CART est un algorithme qui nécessite de nombreux calculs, du fait qu'il choisit la meilleure caractéristique discriminatoire locale à chaque étape du processus [BFOS84] [SC95] [SC97].

J48

L'arbre de décision J48 [Q+92] met en œuvre l'algorithme C4.5 de Quinlan [Qui93] pour générer un arbre C4.5 élagué ou non élagué. C4.5 est une extension de l'algorithme ID3 de Quinlan. Les arbres de décision générés par J48 peuvent être utilisés pour la classification. J48 construit des arbres de décision à partir d'un ensemble de données d'entraînement étiquetées en utilisant le concept d'entropie de Shannon. Il utilise le fait que chaque attribut des données peut être utilisé pour prendre une décision en divisant les données en sous-ensembles plus petits. J48 examine le gain d'information normalisé (différence d'entropie) qui résulte du choix d'un attribut pour la division des données. Pour prendre la décision, l'attribut présentant le gain d'information normalisé le plus élevé est utilisé. Ensuite, l'algorithme revient sur les sous-ensembles. La procédure de division s'arrête si toutes les instances d'un sous-ensemble appartiennent à la même classe. J48 permet de travailler à la fois avec des données discrètes et des données continues. Il permet également de travailler avec des valeurs d'attribut manquante et des attributs ayant des coûts différents. De plus, il offre une option pour élaguer les arbres après leur création. Comparativement à d'autres méthodes de fouille de données, les arbres de décision présentent l'avantage de simplicité de compréhension et d'interprétation.

Les réseaux de neurones

L'inspiration pour les réseaux de neurones provient de la volonté de créer des systèmes artificiels sophistiqués, voire intelligents, capables d'effectuer des opérations semblables à celle que le cerveau humain effectue de manière routinière [Par04]. Un réseau de neurone est un assemblage de neurone formel associé en couches fonctionnant en parallèle. Dans un réseau, chaque sous-groupe fait un traitement indépendant des autres et transmet le résultat de son analyse au sous-groupe suivant. L'information donnée au réseau va donc se propager couche par couche, de la couche d'entrée à la couche de sortie, en passant soit par aucune, une ou plusieurs couches intermédiaires (couches cachées). Excepté par les couches d'entrées et les couches de sorties, chaque neurone dans une couche est connecté à tous les neurones de la couche suivante et de la couche précédente.

Les réseaux de neurones à base radiale (RBF)

Les réseaux de neurones à base radiale sont une classe particulière des réseaux de neurones multicouches. Utilisé pour les problèmes d'apprentissage supervisé [Par04] [CCA96]. En utilisant les réseaux RBF, la formation des réseaux est relativement rapide en raison de la structure simple de ces derniers. Les réseaux RBF peuvent être mis en œuvre dans n'importe quel type de modèle, linéaire ou non linéaire, et dans n'importe quel type de réseau, simple ou multicouche [CCA96]. La conception de ce type de réseau, dans sa forme la plus élémentaire, consiste en trois couches distinctes. La couche d'entrée représente l'ensemble des nœuds sources (unités sensorielles). La deuxième couche consiste en une seule couche cachée de haute dimension. Le rendement (sortie) des nœuds de la couche cachée est déterminé par une fonction d'activation non linéaire (fonction gaussienne). La couche de sortie donne la réponse du réseau aux modèles d'activation appliqués à

la couche d'entrée.

Support Vector Machines (SVM)

Les machines à vecteurs de support, ou Support Vector Machine (SVM), sont des algorithmes d'apprentissage automatique, utilisés en machine learning pour résoudre des problèmes de classification, de régression ou de détection d'anomalie. Les SVM sont considérées comme une généralisation des classificateurs linéaires. Ces machines à vecteurs de support ont été développées dans les années 1990, à partir du concept des informaticiens russes Vladimir Vapnik et Alexey Chervonenkis. Leurs travaux ont vite été adoptés en raison de leur capacité à travailler avec des données de grandes dimensions, leurs garanties théoriques et les bons résultats réalisés en pratique. Les SVM sont appréciées pour leur simplicité d'usage. Dans les SVM, les données sont séparées en plusieurs classes avec une frontière de séparation choisie pour maximiser la distance entre les groupes de données. La difficulté est de déterminer cette frontière optimale. Le problème a été formulé comme un problème d'optimisation. Les points les plus proches de la frontière sont nommés "vecteurs support", d'où le nom de cette famille d'algorithmes [MF06]. Les SVM ont été appliquées avec succès à un certain nombre de problèmes du monde réel, tels que la reconnaissance de caractères et de chiffres manuscrits, la reconnaissance de visages, la catégorisation de textes et la détection d'objets en vision industrielle [CCST98], [CV95], [PV98]. Les SVM trouvent des applications dans l'exploration de données, la bio-informatique, la vision par ordinateur et la reconnaissance des formes [YYH$^+$10].

1.5.2 Études comparatifs des algorithmes de classification

Afin de sélectionner le meilleur classificateur à utiliser parmi les nombreuses méthodes de classification existantes, des études comparatives des différents classificateurs en Data Mining ont été effectuées.

Par exemple dans [SAZ12, IHSM15, AR$^+$11] : les auteurs ont présenté une comparaison entre les différents classificateurs du Data Mining, en utilisant différents ensembles de données sur le cancer du sein. Dans [SAZ12] et [IHSM15] les auteurs présentent une comparaison entre les algorithmes suivants : arbre de décision (J48)[Qui92], Multi-Layer Perception (MLP) [TM18], le modèle Bayesien Naïf (NB) [Bel08a], Sequential Minimal Optimization (SMO) [Pla98], et Instance Based for K-Nearest neighbor (IBK)[ST09] en utilisant trois jeux de données différents sur le cancer du sein (WBC, WDBC et WPBC), les critéres de performance utilisés sont l'exactitude et la matrice de confusion. Ils introduisent aussi un multi classificateur qui est la fusion entre les différents classificateurs. Pour prédire la classe de chaque individu avec ce multi classificateur, les auteurs combinent le résultat de prédiction de chaque classificateur. Dans [IHSM15] les auteurs comparent les approches précédentes en utilisant un autre jeu de données (BCD, Breast Cancer dataset). Dans [AR$^+$11] les auteurs comparent les critères de performances des classificateurs d'apprentissage supervisée tels que la méthode de classification naïve Bayesienne, RBF networks (Réseaux de

neurones à base radiale), Arbre de décisions J48 et Simple CART. Ces classificateurs sont utilisés pour classifier les ensembles de données sur le cancer de sein (WBC, WDBC and Breast tissue). Les méthodes de classification dans [LHM+98b, YH03, LHP01, TCP04, VL08] représentent les algorithmes de classification basés sur les règles. Ils fournissent de bons résultats et permettent de générer des règles d'associations en utilisant l'algorithme Apriori [AS+94]. Ce dernier a été amélioré depuis, et a été appliqué avec succès à de très grands ensembles de données.

Dans [LHM+98b], les auteurs ont introduit la première méthode de classification intéressante basée sur les règles d'association (CBA), en combinant les règles de classification selon un certain ordre, de sorte que les observations puissent être classées en utilisant, les règles les plus fortes, ou bien selon une classe par défaut, de sorte que le taux d'erreur de classification est faible [AAH14, AAT15]. La force des règles est mesurée par la confiance, car elle estime la probabilité de la classe compte tenu des autres caractéristiques de la règle. Leurs résultats sont positifs en ce qui concerne les courbes ROC [1], mais ils produisent généralement des structures assez complexes décrivant le processus de classification.

Dans cette comparaison, les résultats expérimentaux montrent qu'aucune technique de classification parmi les classificateurs n'est meilleure. La classification dépend du type des ensembles de données. Ces méthodes de classifications donnent de bons résultats, mais elles ne sont pas compréhensibles par l'être humain. Dans notre approche, nous utilisons les règles fournis par l'ASI pour avoir des classifications plus compréhensibles.

Dans cette thèse nous avons proposée deux méthodes basées sur les règles avec deux différences principales par rapport à la stratégie CBA. Premièrement, seules les règles de classification à 2 longueurs sont prises en compte (et sont combinées pour l'amélioration de la classification). Ensuite, une nouvelle mesure de qualité est utilisée, l'implifiance qui prend en compte à la fois la confiance des règles et l'effet statistique des prémisses sur les conclusions des règles (l'intensité d'implication).

1.6 Conclusion

Nous avons présenté dans ce chapitre les concepts sur lesquels se basent les contributions de cette thèse. Nous avons expliqué la théorie ASI à travers sa définition, sa modélisation de base et les principaux indices développés pour surmonter aux problèmes des autres mesures de qualité. Nous avons également présenté le logiciel (R)CHIC conçu pour l'ASI, constitués de deux arbres et d'un graphe. Finalement nous avons décrit les différentes techniques de classifications en Data Mining que nous avons utilisé pour évaluer les deux classificateurs proposés dans cette thèse. De plus, une étude comparative des classificateurs les plus utilisée et les plus performant est présentée.

1. Les courbes ROC (fonctions d'efficacité du récepteur) sont un outil important pour évaluer les performances d'un modèle de Machine Learning (ML). Elles sont le plus souvent utilisées pour des problèmes de classification binaire dont la sortie est composée de deux classes distinctes.

Chapitre 2

Analyse des notes des étudiants par le logiciel CHIC

Sommaire

2.1 Introduction 34

2.2 Étude des notes des étudiants en informatique de l'université de Bejaia 35

2.2.1 Étude des notes des étudiants licence 2 durant les années scolaires 2010-2011, 2011-2012 ainsi que 2012-2013 37

2.2.2 Interprétation des résultats qui se répètent dans les trois générations . . 40

2.2.3 Étude des notes des étudiants licence 3 durant les années scolaires 2010-2011, 2011-2012 ainsi que 2012-2013 41

2.2.4 Interprétation des résultats qui se répètent dans les trois générations . . 43

2.3 Conclusion 44

2.1 Introduction

Afin de montrer l'intérêt de l'analyse statistique implicative et sa capacité d'analyser les différents modules de formation d'un système étudié, nous l'avons appliqué aux notes des étudiants de l'université de Bejaia à travers le logiciel CHIC. Notre but est d'analyser les relations entre les modules étudiés par les étudiants proposés en fonction des résultats de ces derniers à travers les différentes promotions. CHIC a été déjà utilisé dans le cadre de la hiérarchisation de compétences tel que dans [Mal10] [MC11] [Dem04] [Ale71] [CD89] [Bay00] mais la plupart de ces études sont purement théorique et peu d'ouvrages en parlent. De plus, elles sont limitées à des populations bien spécifiées [Dem04]. Les auteurs de ces travaux se sont généralement concentrés sur l'amélioration de la mesure d'évaluation des compétences et ils n'ont pas abordé les relations entre les modules étudiés.

Découvrir les liens entre les modules proposés par une formation peut aider les étudiants à choisir un axe de recherche ou une spécialité dans le futur selon les liens existants entre les modules qu'ils maitrisent dans les années précédentes et ceux présents dans la spécialité. En effet, l'identification d'un module maîtrisé par des étudiants qui est nécessaire à la maîtrise d'un autre module (et cela pendant plusieurs années) peut aider les établissements à identifier les causes des échecs des étudiants (peut être la transmission de connaissances entre l'enseignant et les étudiants ne passe pas) et à proposer une solution alternative. Dans ce cadre, le logiciel CHIC, utilisant l'analyse statistique implicative [Cou01] [GRG09] [GKB01], offre l'opportunité de mettre en évidences ces relations. Nous souhaitons illustrer à travers des exemples réalisés sur les notes des étudiants en informatique de l'université de Bejaia qu'il est possible de représenter certaines implications entre les différents modules étudiés. Ces implications sont construites avec la méthode de l'analyse statistique implicative. Cette dernière permet, par exemple, de visualiser par l'intermédiaire de graphes les interactions entre les modules. Ainsi, les enseignants peuvent comprendre si les liens qu'ils observent à travers le graphe semble logique ou non aux vues de différents paramètres (liens entre les modules, comportement des étudiants, interaction des enseignants, ...) afin de déduire les relations entre les notions. Nous avons opté dans ce travail pour la théorie classique, vu le nombre d'étudiants utilisés.

Dans la suite de ce chapitre nous étudions les notes des étudiants en informatique de l'université de Bejaîa pour les deux niveaux licence2 et licence3 durant les années scolaires 2010-2011, 2011-2012 et 2012-2013. Finalement, nous présentons nos expérimentations et les différents résultats obtenus avec quelques interprétations.

2.2 Étude des notes des étudiants en informatique de l'université de Bejaia

Pour analyser les notes des étudiants, nous avons exploité la représentation graphique fournit par le logiciel CHIC. Plus exactement le graphe implicatif. A travers ce graphe les implications précédemment calculées dans «transactions.out » dont l'intensité d'implication est supérieure à un certain seuil seront représentées. Par défaut toutes les variables impliquées dans le graphe sont représentées dans la zone de travail.

Cette analyse représente un exemple concret d'utilisation de CHIC sur les notes des étudiants en informatique de l'université de Bejaïa. Nous nous sommes concentrés sur les niveaux : licence 2 et 3 durant les années scolaires 2010-2011, 2011-2012 et 2012-2013. Pour chaque année scolaire dans un niveau donné nous obtenons un graphe d'implication qui sera comparé aux graphes réalisés dans d'autres années concernant le même niveau. Les modules enseignés en licence 2 sont les suivants :

- Structure de données (StrD).
- Analyse Numérique (ANum).
- Systèmes d'information (SI).
- Architecture (Arch).
- Probabilité et statistiques (PS).
- Logique Mathématique (LogMat).
- Traitement du signal (TSig).
- Anglais semestre 1 (AngS1).
- Algorithme et structure de données (AlStrD).
- Base de Données (BDD).
- Systéme d'Exploitation (SE).
- Théorie des langage (THL).
- Programmation Linéaire (PL).
- Génie Logiciel (GL).
- Anglais semestre 2 (AngS2).

Les modules enseignés en licence 3 sont les suivants :

- Compilation (Compil).
- Interface Homme Machine (IHM).
- Théorie des graphes (ThGraph).
- Programmation Logique (ProgLog).
- Sécurité (Securite).
- Système d'Exploitation (SE).
- Système distribué (SD).
- Cryptographie (Crypto).

	Archi p	STRD1 p	SI p	ANum p	ProStat p
0809TMI02	11,25	9,88	11,5	12,38	11,5
09MI0034	9,38	11,38	7,33	11	10,33
09MI0590	10,63	9,38	8,67	13,75	8,17
09MI0061	11,38	14,69	10,5	12,75	12,17

TABLEAU 2.1 – Exemple de données de type .csv

— Réseaux (Reseaux).

— Projet de fin de cycle (Projet).

Pour lancer l'exécution de ces données avec CHIC nous avons créé des fichiers de type «.csv» (un extrait est montré dans le tableau1) qui contiennent en lignes les matricules des étudiants et en colonnes les modules suivis par ces derniers et la variable « p » qui suit chaque intitulé de module. Ceci précise que les variables (notes de modules) sont à partitionner en un nombre fixe d'intervalles. Le nombre d'intervalles est choisi par l'utilisateur et ensuite l'algorithme des nuées dynamiques [Did71] constitue automatiquement les intervalles qui ont des limites distinctes. Cet algorithme a la particularité de construire des intervalles en minimisant l'inertie de chaque intervalle. Ensuite, un intervalle est représenté par une variable binaire et un individu a la valeur 1 s'il appartient à cet intervalle et 0 sinon. En utilisant une telle décomposition, un individu appartient à un seul intervalle.

L'analyse des résultats a été effectuée en traçant le graphe implicatif correspondant aux données introduites. Nous avons choisi le partitionnement par défaut, ce qui veut dire que CHIC partitionne chaque intervalle de note en trois sous intervalles identifiant les étudiants avec des résultats faibles, moyens et bons. Par exemple, le module STRD1 a été partitionné selon les intervalles suivants : STRD11 de 5.94 à 10.38, STRD12 de 10.5 à 13.38, STRD13 de 13.5 à 18.38. STRD11 reflète les étudiants qui sont faibles en STRD1, STRD12 les étudiants qui sont moyens et STRD13 les étudiant qui sont bons en STRD1.

En premier lieu nous commençons par présenter un petit exemple de graphe implicatif dans lequel nous avons fait apparaître les différents liens ayant des intensités d'implication suffisamment forte pour pouvoir en dégager des règles d'implication ayant du sens. Pour cela nous avant utilisé trois seuils comme le montre la figure suivante :

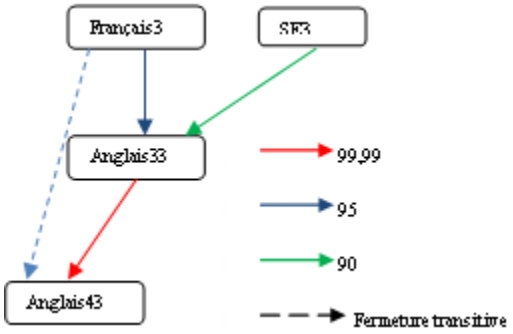

FIGURE 2.1 – Exemple de graphe implicatif.

Ce graphe implicatif permet de visualiser les liens existants entre les modules présents sur la figure. La maîtrise de la langue française implique la maîtrise de la langue anglaise étudiée en premier semestre, impliquant elle-même la maîtrise de la langue anglaise étudié en deuxième semestre, avec des seuils différents. Aussi nous remarquons une implication transitive entre Francais3 et Anglais43, elle est représentée par l'intensité d'implication entre Français3 et Anglais43 qui est plus élevée que la neutralité ($\geq 0,5$) [Gra96]. Au final, nous présentons des graphes implicatifs « épuré » où les seules fermetures transitives retenues sont celles ayant une intensité élevée.

2.2.1 Étude des notes des étudiants licence 2 durant les années scolaires 2010-2011, 2011-2012 ainsi que 2012-2013

Dans ce qui suit nous présentons les graphes implicatifs correspondant aux notes des étudiants en licence 2 durant les années scolaires 2010-2011 (Figure 2.2), 2011-2012 (Figure 2.3) et 2012-2013 (Figure 2.4). Ensuite, nous extrayons les relations d'implications les plus significatives c'est-à-dire celles qui se répètent durant les trois années puis nous présentons leur signification dans le prochain paragraphe.

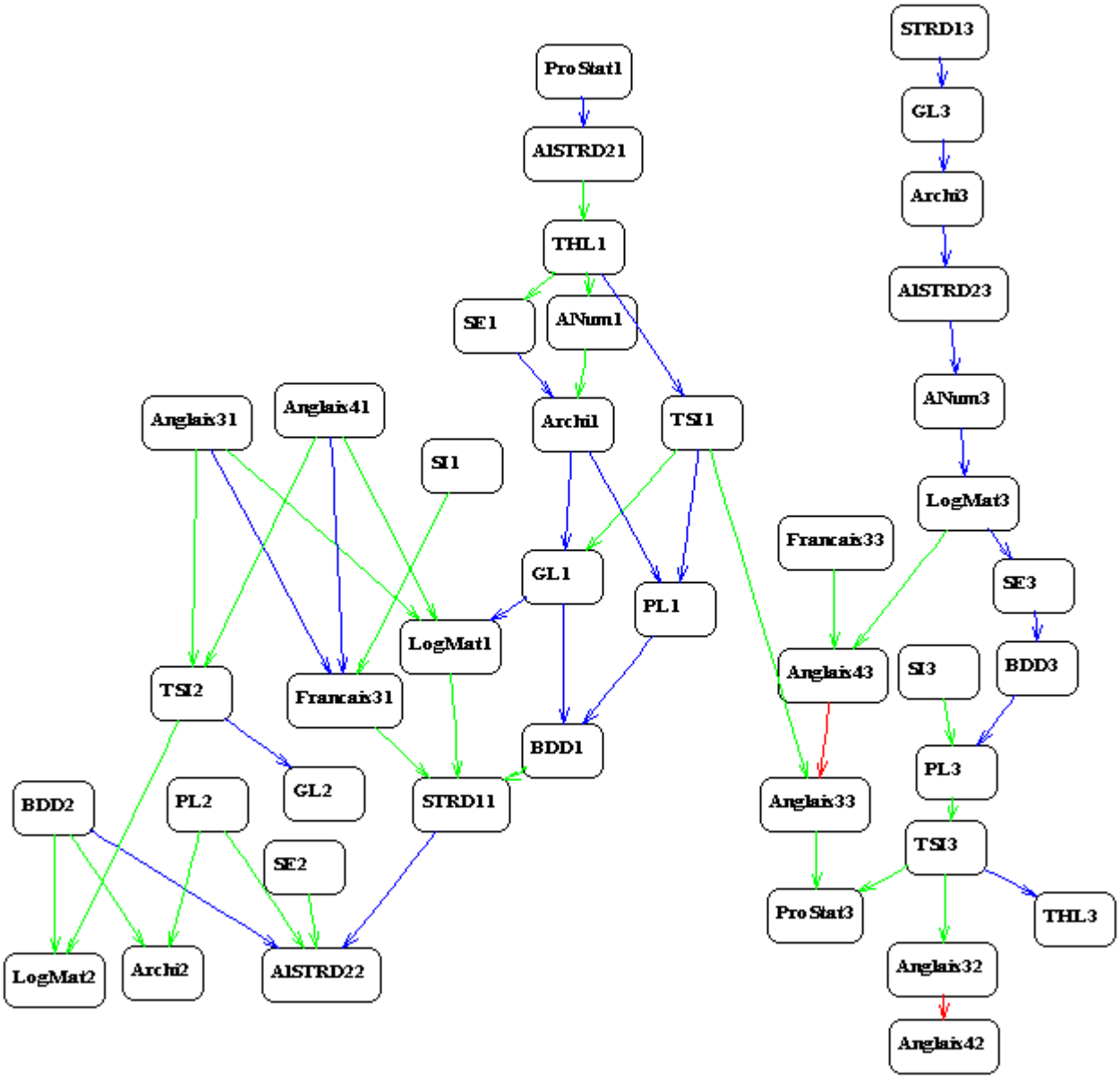

FIGURE 2.2 – Graphe implicatif Licence 2 2010-2011.

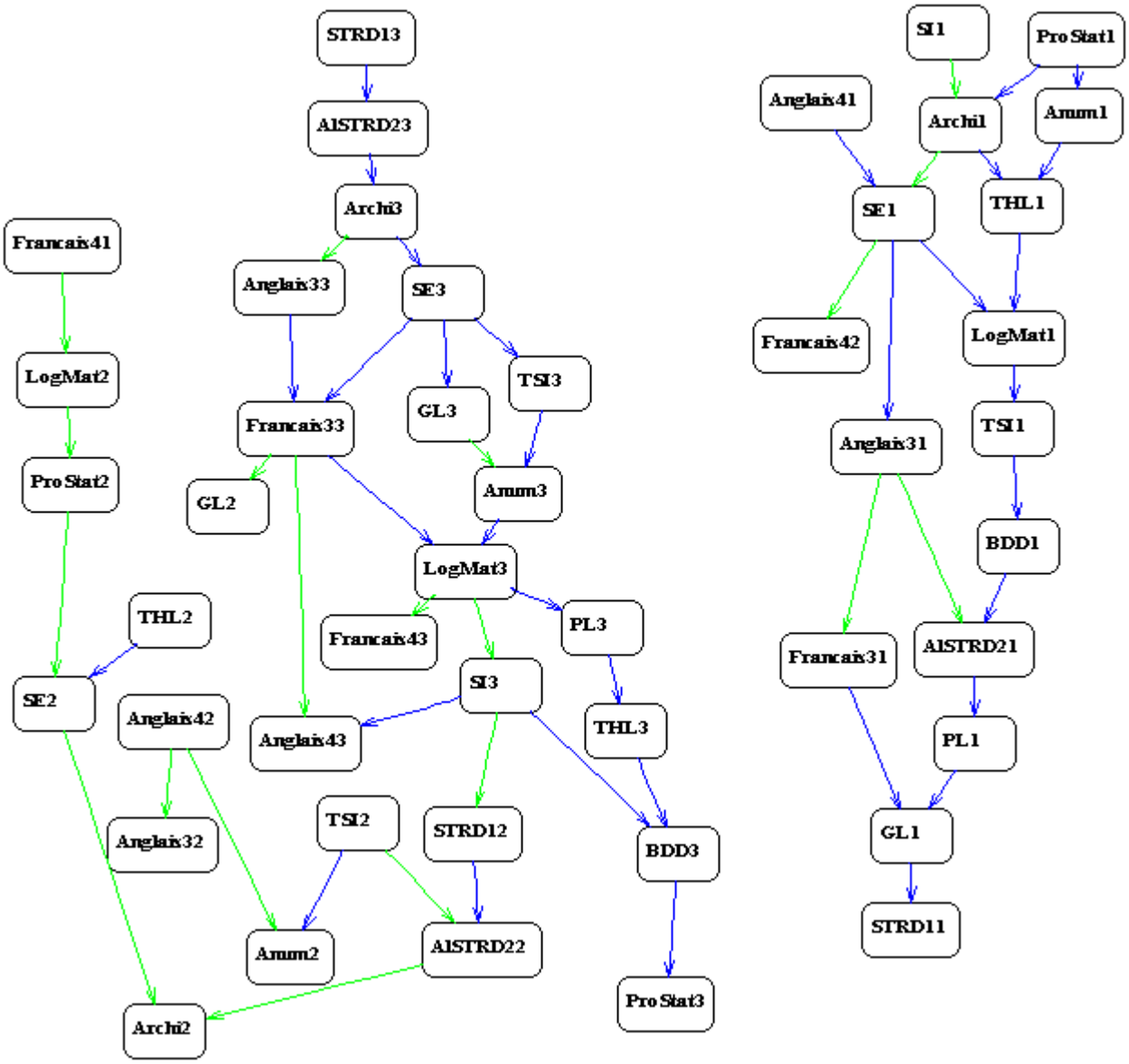

FIGURE 2.3 – Graphe implicatif Licence 2 2011-2012.

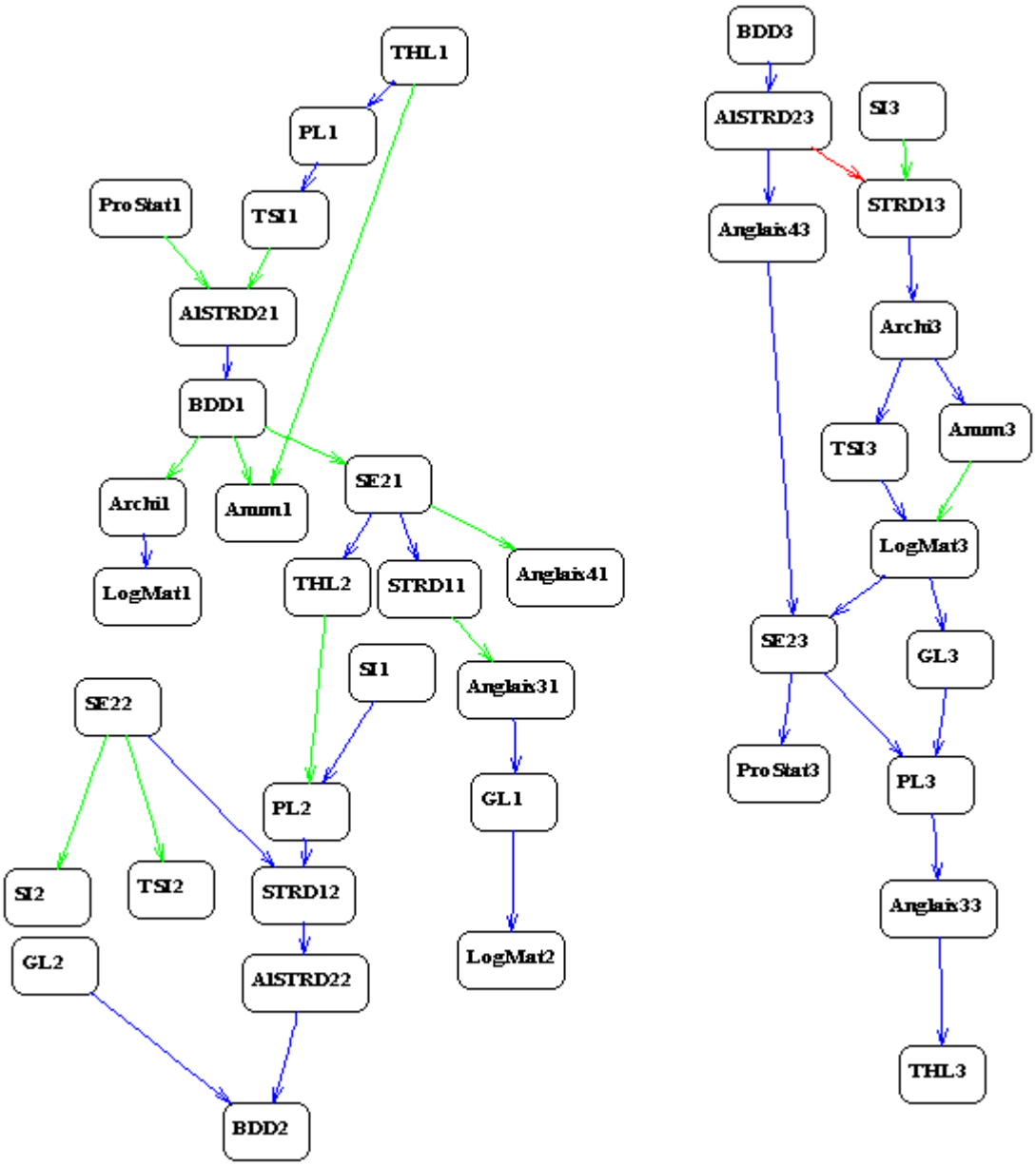

FIGURE 2.4 – Graphe implicatif Licence 2 2012-2013.

2.2.2 Interprétation des résultats qui se répètent dans les trois générations

Premièrement nous avons remarqué qu'on observe presque les mêmes implications qui se répètent pour les trois générations d'étudiants. Nous remarquons clairement qu'un module réussi par les bons élèves entraine d'autres modules réussis par les bons élèves. Il en est de même pour les modules réussis (ou échoués) par les élèves moyens ou faibles qui respectivement impliquent d'autres modules réussis (ou échoués) par des élèves moyens ou faibles. Pour l'implication $Anum3 \rightarrow LogMat3$ qui se répètent en 2010, 2011 et en 2012, les deux modules Analyse Numérique (Anum) et Logique Mathématique (LogMat) se basent sur les mathématiques.

Pour les implications $STDR13 \rightarrow AlSTDR23$ (2011) $STDR12 \rightarrow AlSTDR22$ (2011, 2012) $STRD11 \rightarrow AlSTDR22$ (2010), le module Algorithme et Structures de Données 2 (AlSTRD 2) assuré au deuxième semestre est la suite du module Structure de Données1 (STRD 1) qui se fait en premier semestre. Pour les implications $Anglais31 \rightarrow Francais31$ (2010) $Anglais33 \rightarrow Francais33$ (2011) $Anglais41 \rightarrow Francais31$ (2010), les étudiants forts en Anglais le sont aussi en Français et ceux qui sont faibles le sont aussi. Il y a en général un lien logique derrière cela. En général, les gens qui maîtrisent et s'intéressent aux langues peuvent maîtriser d'autres langues. Ceci est visible à travers $Anglais42 \rightarrow Anglais32$ (2011) $Anglais43 \rightarrow Anglais33$ et $Anglais33 \rightarrow (FT)Anglais43$ (2010)$Anglais32 \rightarrow Anglais42$ et $Anglais42 \rightarrow (FT)Anglais32$ (2010) $Anglais31 \rightarrow (FT)Anglais41$ et $Anglais41 \rightarrow (FT)Anglais31$. Le module Anglais4 assuré au deuxième semestre est la suite du module Anglais3 fait au premier semestre. Pour $Francais33 \rightarrow Anglais43$ (2010). Les étudiants qui sont faibles en français le seront malheureusement généralement en anglais. Pour $SE1 \rightarrow Anglais31$ (2011) $SE21 \rightarrow Anglais41$ (2011) $Anglais43 \rightarrow SE23$ (2012), les étudiants qui sont bons en Système d'Exploitation au premier semestre le sont aussi en module Anglais du premier semestre et ceux qui sont bons en module Système d'Exploitation du deuxième semestre le sont aussi en module Anglais du deuxième semestre. Ceci s'explique par le fait que les sujets d'anglais portent sur le module Système d'Exploitation pour l'année 2011. Par contre, pour les autres années, le sujet d'Anglais porte sur l'anglais général, c'est pour cela qu'on ne voit pas l'implication. Pour la règle $GL1 \rightarrow BDD1$ (2010) et $GL2 \rightarrow BDD2$ (2012) on explique cette implication par le fait que les deux modules se basent sur la conception (BDD conception de projets et Gl conception de logiciels). Pour les implications $THL1 \rightarrow SE1$ (2010) $THL2 \rightarrow SE2$(2011) et $SE21 \rightarrow THL2$ (2012), le module Théorie des Langages (THL) contient des graphes (comme les automates d'état fini) et pour Système d'Exploitation (SE) les exemples d'application se basent sur la construction de graphes par exemple algorithme de Banquier, graphe d'allocation, graphe d'attente, etc. Pour les implications $THL1 \rightarrow Anum1$ (2010, 2012) et $Anum1 \rightarrow THL1$(2011), le module de Théorie des Langages (THL) contient la construction des grammaires à partir de langages et ceci utilise la logique qui se base sur les mathématiques tout comme l'Analyse numérique. Pour les implications $BDD2 \rightarrow AlSTDR22$ (2010), $BDD1 \rightarrow AlSTDR21$ (2011), $BDD3 \rightarrow AlSTDR23$ (2012) et $BDD1 \rightarrow AlSTDR11$ (2010), dans les Bases de Données les étudiants étudiés souvent le langage SQL donc c'est assez proche de l'algorithmique du module Algorithme et Structure de Données.

2.2.3 Étude des notes des étudiants licence 3 durant les années scolaires 2010-2011, 2011-2012 ainsi que 2012-2013

Comme précédemment, nous présentons les graphes implicatifs correspondant aux notes des étudiants en licence 3 durant les années scolaires 2010-2011 (Figure 2.5, 2011-2012 ((Figure 2.6 ainsi que 2012-2013 (Figure 2.7). Ensuite nous essayons d'extraire les relations d'implications les

plus significatif, c'est-à-dire celles qui se répètent durant les trois années.

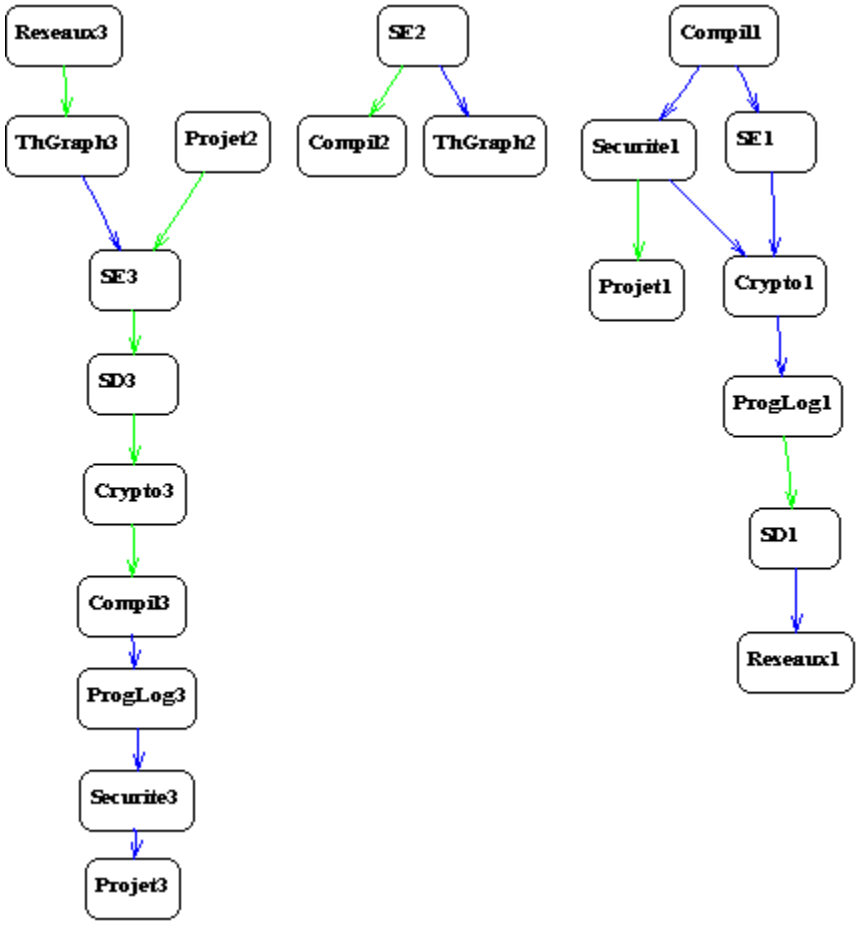

FIGURE 2.5 – Graphe implicatif Licence3 2010-2011.

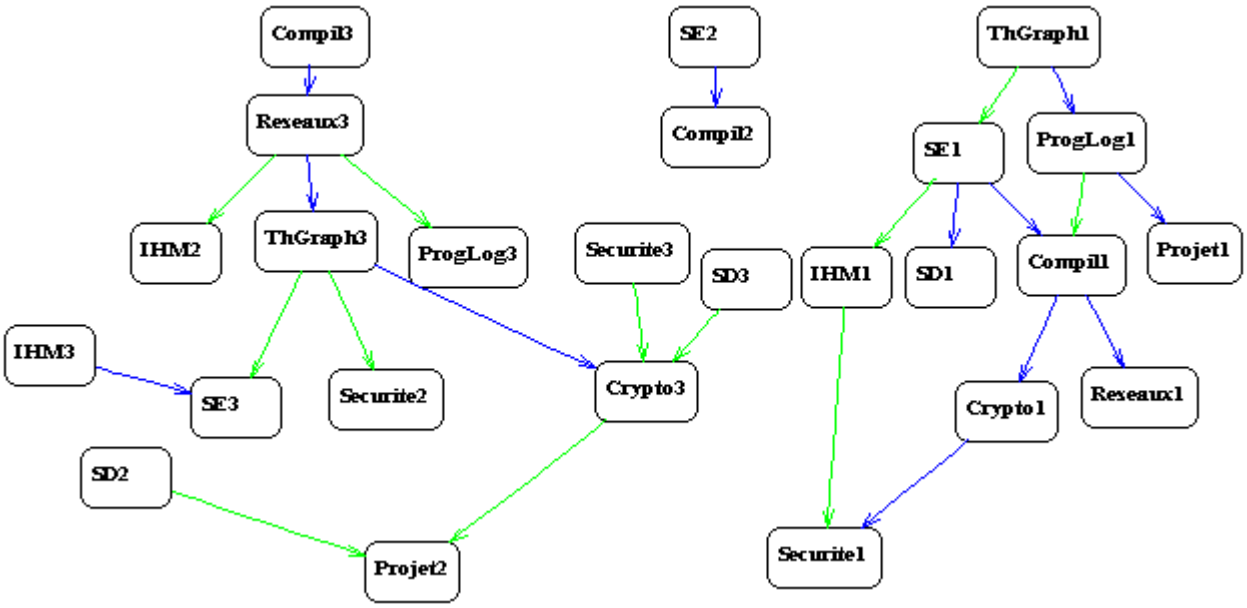

FIGURE 2.6 – Graphe implicatif Licence3 2011-2012.

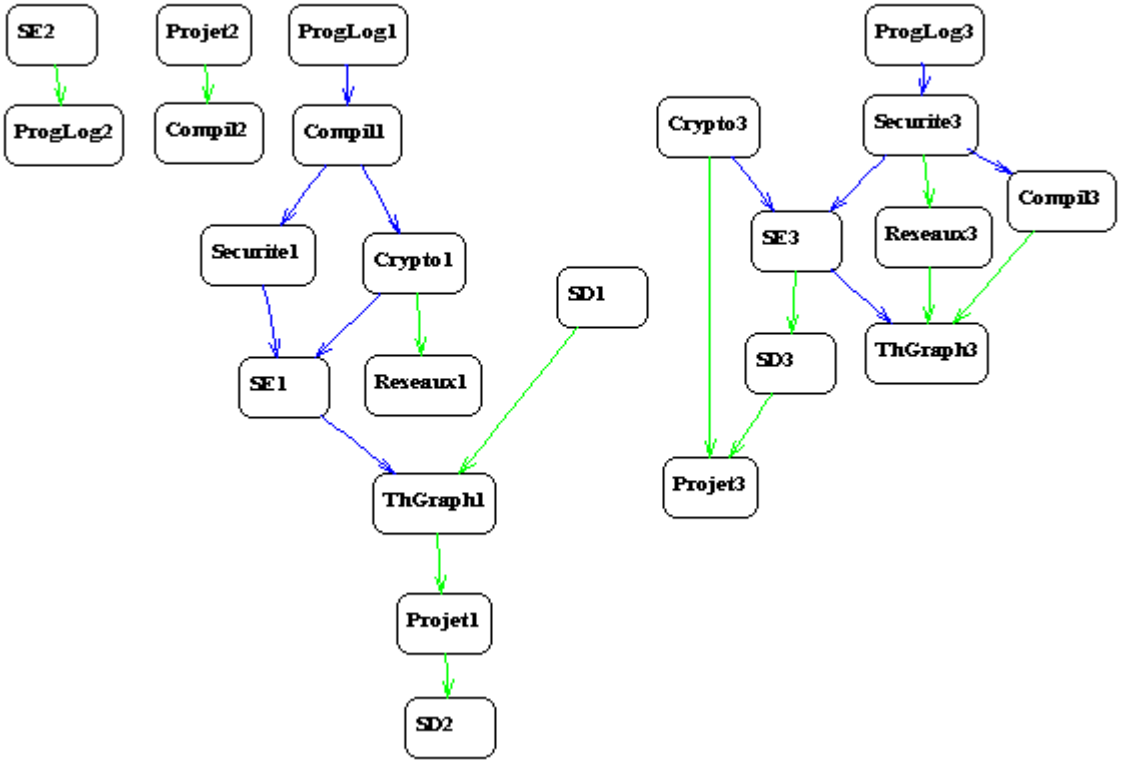

FIGURE 2.7 – Graphe implicatif Licence3 2012-2013.

2.2.4 Interprétation des résultats qui se répètent dans les trois générations

Comme pour les Licence 2 nous remarquons qu'on a presque les mêmes implications qui se répètent pour les trois générations. Nous voyons comme précédemment que les modules réussis par les bons élèves impliquent d'autres modules également réussi par les bons élèves. Il en va de même pour des implications entre modules validés ou échoués par des élèves moyens ou faibles. Pour l'implication $SE3 \rightarrow SD3$ (2010, 2012) et $SE1 \rightarrow SD1$ (2011) cela se justifie par le fait que le programme de module Système Distribué est un chapitre dans le module Système d'Exploitation mais avec plus de détail. Pour $ProgLog1 \rightarrow Compil1$ (2011, 2012)$Compil3?ProgLog3$ (2010), les deux modules Programmation Logique et Compilation sont assurés par le même enseignant. Pour $Securite1 \rightarrow Crypto1$ (2010) $Securite3 \rightarrow Crypto3$ (2011)$Crypto1 \rightarrow Securite1$ (2011), le module Cryptographie est un chapitre dans le module sécurité avec plus de détail.

Pour $SE2 \rightarrow ThGraph2$ (2010) $SE1 \rightarrow ThGraph1$ (2011) $SE3 \rightarrow ThGraph3$ (2011) $ThGraph3 \rightarrow SE3$ (2010, 2012), le module Système d'Exploitation (SE) contient plusieurs algorithmes et ceux-ci sont illustrés par la construction de graphes donc un étudiant qui maîtrise ce module va sûrement maîtriser le module Théorie des Graphes.

2.3 Conclusion

Dans le but de montrer l'intérêt de l'Analyse Statistique Implicative (ASI) nous avons étudié les notes des étudiants en Informatique de l'Université de Bejaïa par l'intermédiaire du logiciel CHIC. Les résultats obtenus montrent que les implications se répètent sur les trois générations et elles sont intéressantes du fait que les bons impliquent les bons, les moyens impliquent les moyens et les faibles impliquent les faibles. Nous avons aussi remarqué des résultats d'implication entre les modules qui sont réellement liés par leur contenu ou par l'enseignant qui les assure ce qui affirme l'efficacité de la méthode d'ASI ainsi que le logiciel CHIC.

Cette étude a aussi comme but de convaincre les administrations d'utiliser l'ASI pour fournir une éducation scientifique pour les futurs étudiants. Elle permet d'aider les étudiants à choisir une spécialité donnée. Elle donne une séquence des concepts ou un module doit être enseigné avant un autre ; par exemple l'implication qui se produit entre la matière Génie logiciel et base de données. Elle permet aussi de déterminer la cause des situations d'échec. Par exemple si une situation survient plusieurs fois, a l'exception d'une année ; cela peut indiquer à l'administration que la situation pourrait être causée par exemple par le professeur.

Ces résultats montrent l'efficacité de la méthode ASI ainsi que le logiciel CHIC qui a été conçu il y a 15 années en C++. Le logiciel R est actuellement le plus utilisé par les statisticien et regroupe de très nombreuses méthodes à disposition de la communauté scientifique, d'où notre objectif pour développer les méthodes de l'ASI dans R.

Chapitre 3

Ajout de la Confiance au graphe Implicatif

Sommaire

3.1 Introduction 46
3.2 Motivations 47
3.3 L'approche proposée et son application sur des données issues des échocardiographie de stress 49
3.3.1 Résultats obtenus sans spécifier un seuil de confiance 53
3.3.2 Résultats obtenus en utilisant un seuil de confiance égal à 80 54
3.3.3 Résultats obtenus en utilisant un seuil de confiance égale à 70 55
3.3.4 Résultats obtenus avec un seuil de confiance égal à 65 55
3.4 Conclusion 56

3.1 Introduction

L'analyse statistique implicative porte sur les mesures statistiques de qualité des règles d'associations, elle se distingue des autres méthodes statistiques qui permettent de générer des règles d'association, par le fait qu'elle utilise une mesure non linéaire qui satisfait des critères importants, [Cou07, Cou08]. Cette dernière basée sur le calcul de l'intensité d'implication permet d'extraire les règles les plus étonnantes sans prendre en considération le nombre de contribution des individus à chaque règle. Certaines personnes ne sont pas satisfaites par l'ASI car la valeur de confiance de chaque règle n'est pas connue, en effet, seul le degré d'étonnement de la règle est connu. La confiance (probabilité conditionnelle) représente la fonction la plus classique servant à la confirmation inductive d'une règle. Elle fonde la stratégie utilisée dans les principaux algorithmes d'extraction. Ces algorithmes prennent en considération le support et le seuil de confiance pour sélectionner les règles. Ces algorithmes sélectionnent l'ensemble des règles dont le support et la confiance dépassent les seuils de support et de confiance préalablement fixés. Ce type d'algorithme, exhaustif et déterministe [Fre00], produit des règles trop nombreuses dont l'intérêt n'est pas toujours assuré [Gra00].

Dans cette contribution nous proposons de rassembler les deux informations : intensité d'implication et confiance dans le graphe d'implication implémenté dans RCHIC, pour avoir en même temps les règles les plus étonnantes et pour lesquelles le pourcentage de participation est très élevé. Pour montrer l'intérêt de l'ajout de la mesure de confiance nous avons travaillé sur des données conçues à partir des échocardiographies de stress. L'utilisation de ces deux mesures aidera les médecins à sélectionner les règles qui sont fortes en intensité d'implication et en confiance. Le choix d'un seuil de confiance élevé permet d'extraire les causes qui induisent avec une grande probabilité à avoir un état de stress. En diminuant le seuil de confiance cette probabilité d'avoir un état de stress diminue, de ce fait ce seuil permet d'avoir un graphe plus lisible.

Dans cette contribution nous proposons de rassembler les deux informations : intensité d'implication et confiance dans le graphe d'implication implémenté dans RCHIC [1], pour avoir en même temps les règles les plus étonnantes et dont le pourcentage de participation est très élevé. Pour montrer l'intérêt de l'ajout de la mesure de confiance nous avons travaillé sur des données conçues à partir des échocardiographies de stress. L'étude de ces derniers en ajoutant cette nouvelle mesure nous a permis d'identifier les causes qui conduisent à un état de stress. L'utilisation de ces deux mesures aidera les médecins à sélectionner les règles qui sont fortes en intensité d'implication et en confiance. Le choix d'un seuil de confiance élevé permet d'extraire les causes qui induisent avec une grande probabilité à avoir un état de stress. En diminuant le seuil de confiance, cette probabilité d'avoir un état de stress diminue, de ce fait ce seuil permet d'avoir un graphe plus lisible.

Dans la suite de ce chapitre nous allons calculer l'intensité d'implication et la confiance pour

1. RCHIC est la version écrite dans R du logiciel CHIC (elle est encore en cours de développement)

des données expérimentales. A travers les résultats obtenus, nous montrons comment l'expert peut utiliser ces deux informations ensemble pour extraire les règles qui l'intéressent le plus. Par la suite l'approche sera expérimentée par des données issues d'une étude sur l'échocardiographie de stress des malades. Nous présentons ces expérimentations et les différents résultats obtenus avec des interprétations. Enfin nous terminons par une conclusion.

3.2 Motivations

Dans le but d'extraire la connaissance à partir des grandes bases de données, plusieurs algorithmes ont été conçus. Parmi, les algorithmes de type apriori, ces derniers se basent sur la recherche des règles d'association intéressantes. L'inconvénient de ces algorithmes est qu'ils produisent beaucoup de règles qui ne sont pas toutes intéressantes, et ils ignorent d'autres règles intéressantes [LT04]. Pour pallier ces problèmes une autre méthode qui permet de structurer les données en individus et variables a été développée. Il s'agit de l'analyse statistique implicative. La méthode implicative se développe au fil de ces problèmes rencontrés et de ces questions posées. Son objectif majeur vise la structuration de données croisant individus et variables. Dans cette contribution nous utilisons le graphe d'implication mis en œuvre dans RCHIC qui permet d'extraire les implications les plus étonnantes basées sur le calcul de l'intensité d'implication. Le graphe d'implication permet de donner toutes les implications étonnantes.

La question qui se pose est : est-ce que toutes les règles obtenues avec l'ASI avec une intensité d'implication forte sont intéressantes ? La réponse à cette question dépend du domaine étudié et revient en principe à l'expert du domaine. Dans certains cas le nombre de règles fortes en termes d'intensité d'implication est très élevé et le graphe devient moins lisible. Ceci nous a motivé à intégrer une autre mesure qui permet de classer les règles selon leur importance. Il s'agit de calculer la valeur de confiance pour chaque règle. RCHIC calcule cette valeur et l'enregistre dans un fichier appelé transaction.out. La récupération de cette valeur pour chaque règle et son ajout au graphe d'implication permet de déterminer les règles qui sont fortes en termes d'intensité d'implication et qui ont une confiance élevée. Ces dernières pourront être considérées comme les plus intéressantes. Pour cela nous avons effectué une expérimentation sur des données avec des valeurs représentatives, ensuite nous avons calculé l'indice d'implication et la confiance. La variable n correspond au nombre d'individus dans la population totale, les variables na et nb correspondent respectivement aux nombres d'apparition des propriétés a et b et la variable $nabb$ correspond au nombre d'apparition de a et de $non\ b$.

Dans (Tableau 3.1), nous avons initialement pris un échantillon de n variables parmi lesquelles les variables na,nb, $nabb$ apparaissent respectivement 20, 40 et 10 fois. La valeur de confiance calculée est égale à 50% et l'intensité d'implication à 71,81, par la suite nous avons fait varier la taille de l'échantillon. Nous remarquons que la confiance ne dépend pas de la taille de l'échantillon, elle est restée constante. Ceci s'explique par le fait que la confiance de la règle $a => b$ est égale à la

N	Na	nb	Nabb	confiance	Intensité d'implication
100	20	40	10	50%	71,81
150	20	40	10	50%	88,84
200	20	40	10	50%	93,31
250	20	40	10	50%	95,14
400	20	40	10	50%	97,03

Tableau 3.1 – Variation de la taille de l'échantillon n

N	Na	nb	Nabb	confiance	Intensité d'implication
400	20	40	2	90%	99,99
400	20	40	4	80%	99,95
400	20	40	6	70%	99,76
400	20	40	8	60%	99,07
400	20	40	10	50%	97,03
400	20	40	12	40%	92,13
400	20	40	14	30%	82,71
400	20	40	16	20%	68,13
400	20	40	18	10%	50,00

Tableau 3.2 – Variation de la variable *nabb*

probabilité conditionnelle d'avoir b sachant a, cette dernière est égale au nombre d'apparition de la variable a union b divisé par le nombre d'apparition de la variable a. Ainsi, dans (Tableau 3.1), la confiance indépendante de la taille de l'échantillon. Par contre l'intensité d'implication augmente avec la taille de l'échantillon et montre ainsi la croissance de la surprise lorsque n croit.

Dans (Tableau 3.2), nous avons fixé les valeurs n, na et nb et nous avons fait varier le nombre d'apparitions de a et non b ($nabb$). Nous remarquons que les valeurs de confiance et d'intensité d'implication sont très fortes lorsque le nombre de $nabb$ est faible et elles diminuent avec sa diminution. Avec l'augmentation du nombre de $nabb$ les valeurs de confiance et d'intensité d'implication diminue considérablement, cela parait logique car plus on a de contre-exemples d'une règle plus le nombre d'exemples qui vérifient la règle diminue ce qui implique la diminution de la valeur de confiance et d'intensité d'implication.

Nous avons fait varier la variable na dans (Tableau 3.3) et la variable nb dans (Tableau 3.4). En augmentant la valeur de na, les valeurs de confiance et d'intensité d'implication augmentent.

N	Na	nb	Nabb	confiance	Intensité d'implication
400	20	40	10	50%	97,03
400	25	40	10	60%	99,58
400	28	40	10	64,29%	99,88
400	35	40	10	71,43%	99,99

Tableau 3.3 – Variation de la variable *na*

N	Na	nb	Nabb	confiance	Intensité d'implication
400	20	40	10	50%	97,03
400	20	50	10	50%	96,35
400	20	55	10	50%	95,95
400	20	100	10	50%	90,16

Tableau 3.4 – Variation de la variable *nb*

Par contre en augmentant la variable *nb* la confiance reste constante et l'intensité d'implication diminue.

À partir des résultats présentés dans ces tableaux, nous constatons que la modification de la taille de l'échantillon et du nombre d'apparition de la variable en conclusion d'une règle d'implication n'ont aucune influence sur la valeur de confiance par contre en augmentant la taille de l'échantillon l'intensité d'implication augmente. En augmentant le nombre d'apparition de la variable en conclusion d'une règle l'intensité diminue. Dans les cas où la valeur de confiance est faible contrairement à la valeur de l'intensité d'implication qui est très élevée, notre approche offre la possibilité à l'expert de prendre ces implications ou de les ignorer et cela en choisissant un seuil de confiance. Ces dernières sont supposées moins importantes que les règles qui ont une valeur de confiance et d'intensité d'implication élevée qui sont supposées les plus importantes.

Dans la suite nous allons voir l'impact d'élimination des règles d'implication qui ont une valeur de confiance faible et cela à travers des données issues des échocardiographies de stress.

3.3 L'approche proposée et son application sur des données issues des échocardiographie de stress

Notre approche consiste à combiner les deux mesures, l'intensité d'implication et la confiance. Pour cela dans RCHIC nous avons rajouté la possibilité d'afficher les valeurs de la confiance pour chaque règle en insérant un bouton « confidence » dans le menu du graphe d'implication. Pour sélectionner que les règles considérées comme intéressantes, une fenêtre a été implémenté pour choisir la valeur de confiance désirée.

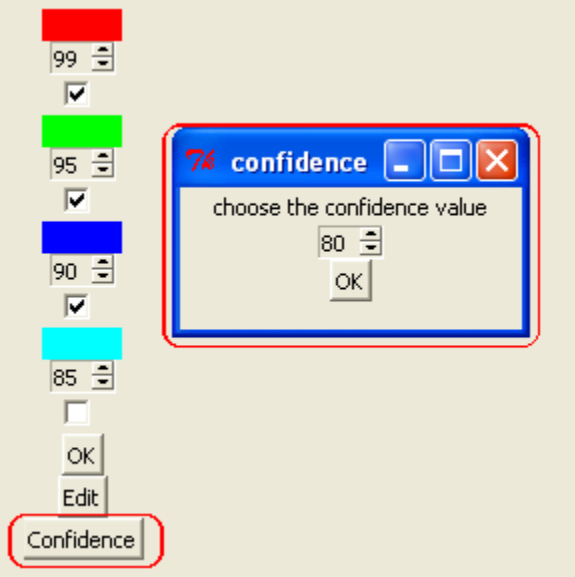

FIGURE 3.1 – Fenêtre permettant le choix de la valeur de confiance.

La suite de cette section présente un exemple concret d'utilisation de RCHIC pour étudier les causes qui conduisent à un état de stress. Nous avons récolté des informations à partir d'une base de données sur l'analyse des données issues des échocardiographie de stress obtenue à l'Université de VANDERBILT par Frank Harrell [Har15]. Il a évalué les échocardiographies de stress de 558 patients jusqu'au 24 février 2015. Le jeu de données contient 31 variables.

L'échocardiographie ou bien échographie cardiaque est un test de diagnostic qui utilise des ondes ultrasonores pour créer une image du muscle cardiaque qui désigne les dimensions et le volume de cœur. Il est ainsi possible de voir la taille, la forme, et le mouvement des soupapes et cavités cardiaques ainsi que la circulation du sang à travers le cœur. L'échocardiographie peut montrer des anomalies du mauvais fonctionnement des valves cardiaques ou des dommages aux tissus du cœur à cause d'une crise cardiaque passé. Avant de présenter les résultats obtenus nous définissons les variables utilisées.

Le jeu de données utilisé comporte les variables suivantes :

— bhr : Fréquence cardiaque de base
— basebp : Tension de base
— basepb : Le produit Bhr* basebp
— pkhr : Le sommet de la fréquence cardiaque
— sbp : La pression systolique [2]
— dp : Le produit pkhr*sbp
— dose : Dose de dobutamine [3] donnée
— mbp : Pression artérielle maximale

2. Correspond à la pression artérielle mesurée lors de la phase de la contraction du c ?ur.
3. Dobutamine est un médicament utilisé pour l'augmentation de la contraction cardiaque, notamment en cas d'insuffisance cardiaque grave

— dpmaxdo : Doubler la dose du dobutamine au maximum dans le produit
— dobdose : La dose du dobutamine lorsque le produit double en maximum
— gender : Le sexe, masculin ou bien féminin
— baseEF : La fraction d'éjection cardiaque initiale
— dobEF : Fraction d'éjection sur dobutamine
— chestpain : Une douleur dans la poitrine
— restwma : Repos à anomalie de mouvement de la paroi sur échocardiogramme
— posSE : Le stress positif échocardiogramme
— newMI : Nouvelle crise cardiaque
— newPTCA : Angioplastie [4] récente
— hxofHT : Le malade est sujet à une hypertension
— hxofDM : Le malade est sujet au diabète
— hxofCig : Si le malade fume : il est soit non fumeur, fumeur modéré ou bien un grand fumeur.
— hxofMI : Vérifier si le malade aune crise cardiaque aiguë et grave.
— hxofPTCA : Vérifier si le malade a subit la chirurgie d'angioplastie
— ecg : Diagnostic de base de l'électrocardiogramme [5], permet de déduire trois cas : un cas normal, désigné par la variable acg.Normal, un cas indéterminé désigné par la variable ecg.equivocal (on ne peut pas confirmer s'il y a un trouble cardiaque ou pas) et un cas de crise cardiaque désigné par la variable ecgMI.

Pour lancer l'analyse de ces données sous RCHIC nous avons créé un fichier de type «.csv» (un extrait est montré dans (Figure 3.2) qui contient en ligne les identifiants des malades et en colonne les variables. Certaines variables sont suivies par la lettre « p ». Ceci précise que les variables sont à partitionner en un nombre fixe d'intervalles. Ensuite l'algorithme des nuées dynamiques [Did71] constitue automatiquement les intervalles qui ont des limites distinctes. Nous avons choisi le partitionnement par défaut ce qui veut dire que RCHIC partitionne chaque intervalle en trois sous intervalles. Par exemple, la fréquence cardiaque de base bhr a été partitionnée en trois intervalles : élevée, moyenne et faible. Un intervalle est représenté par une variable binaire et un individu a la valeur 1 s'il appartient à cet intervalle et 0 sinon. En utilisant une telle décomposition, un individu appartient à un seul intervalle. Pour les variables qui contiennent plus d'un sens, on a attribué une

4. Angioplastie : une technique chirurgicale pour rétablir la circulation normale du sang dans une artère rétrécie ou bloquée par l'athérosclérose, soit par l'insertion d'un ballonnet dans la section rétrécie et le gonfler soit en utilisant un faisceau laser

5. L'électrocardiographie (ECG) désigne l'examen permettant l'enregistrement du rythme cardiaque. L'ECG consiste à étudier précisément l'activité du cœur, grâce à des électrodes posées sur la poitrine, les poignets et les chevilles. Cette activité est mesurée en plusieurs points du cœur, appelés dérivations, et elle est enregistrée sous la forme d'une courbe pour chacune d'entre elles. L'ECG permet de découvrir des troubles du rythme cardiaque, des troubles de la conduction cardiaque, des signes de souffrance cardiaque. L'ECG peut être utilisée pour identifier une population de patients à faible risque de MI (crise cardiaque). Un ECG normal indique moins de 3% pour le risque de MI et moins de 6% pour le risque de décès dans l'année choisie

variable pour chaque sens par exemple à partir de la variable sexe on a créé deux variables « male » et « female »

	bhr p	basebp p	basedp p	pkhr p	sbp p	dp p	dose p	maxhr p	pctMphr p	mbp p	dpmaxdo p	dobdose p	age p	mal	female	baseEF p	dc
1	92	103	9476	114	86	9804	40	100	74	121	12100	40	85	1	0	27	
2	62	139	8618	120	158	18960	40	120	82	158	18960	40	73	1	0	39	
3	62	139	8618	120	157	18840	40	120	82	157	18840	40	73	1	0	39	
4	93	118	10974	118	105	12390	30	118	72	105	12390	30	57	0	1	42	
5	89	103	9167	129	173	22317	40	129	69	176	22704	40	34	1	0	45	
6	58	100	5800	123	140	17220	40	123	83	140	17220	40	71	1	0	46	
7	63	120	7560	98	130	12740	40	98	71	130	12740	40	81	0	1	48	
8	86	161	13846	144	157	22608	40	144	111	157	22608	40	90	0	1	50	
9	69	143	9867	115	118	13570	40	113	81	151	17063	40	81	0	1	52	
10	76	105	7980	126	125	15750	40	126	94	125	15750	40	86	1	0	52	
11	105	134	14070	171	182	31122	40	171	108	182	31122	40	61	0	1	52	
12	72	112	8064	127	95	12065	30	125	80	101	12625	20	63	1	0	53	
13	90	120	10800	169	184	31096	40	169	126	184	31096	40	86	1	0	54	
14	81	110	8910	110	130	14300	40	110	58	130	14300	40	29	0	1	55	
15	84	176	14784	110	194	21340	40	110	74	194	21340	40	71	0	1	55	

FIGURE 3.2 – Extrait du jeu de données.

RCHIC calcul la valeur de confiance et d'indice d'implication pour chaque règle, et les enregistre dans un fichier appelé transaction.out. Ce dernier correspondant au jeu de données issu des données des échocardiographies de stress contient 3986 règles. Voici une partie de ce fichier :

hyp -> con	occurrence(	occurren	suppor	confidence	classical index	entropic index
maxhr.3 -> dose.2	128.000000000	125.000000	22.939068	28.90625000000	79.8263922333717	33.8972340076598540
dose.2 -> maxhr.3	125.000000000	128.000000	22.401433	29.59999999999	80.18773198127746	34.8368000960555620
maxhr.3 -> HxofCig.he	128.000000000	122.000000	22.939068	27.34375000000	75.84681957960128	31.9113793722364250
HxofCig.heavy -> maxh	122.000000000	128.000000	21.863795	28.68852459016	76.5288650989532	33.7130189901168580
maxhr.3 -> bhr.3	128.000000000	108.000000	22.939068	38.28125000000	99.1447148844599	46.7913717142461320
bhr.3 -> maxhr.3	108.000000000	128.000000	19.354838	45.37037037037	99.60406208410859	58.2992773010986060
maxhr.3 -> dobEF.1	128.000000000	100.000000	22.939068	16.40625000000	42.4976348876953	18.6547383807354310
dobEF.1 -> maxhr.3	100.000000000	128.000000	17.921146	21.00000000000	41.2589073181152	24.4447615698129790
maxhr.3 -> dp.3	128.000000000	88.0000000	22.939068	39.06250000000	99.7955969069153	48.0348474151621050
dp.3 -> maxhr.3	88.0000000000	128.000000	15.770609	56.81818181818	99.9942285424292	80.2317373407175440
maxhr.3 -> basedp.3	128.000000000	88.0000000	22.939068	27.34375000000	92.31628403067588	32.0774828447578530
basedp.3 -> maxhr.3	88.0000000000	128.000000	15.770609	39.77272727272	96.39814011752605	49.5371237867437630
maxhr.3 -> hxofCABG	128.000000000	88.0000000	22.939068	15.62500000000	49.2839395999084	17.7773868543899950
hxofCABG -> maxhr.3	88.0000000000	128.000000	15.770609	22.72727272727	49.0971505641937	26.7144851954282670
maxhr.3 -> dpmaxdo.3	128.000000000	84.0000000	22.939068	38.28125000000	99.7822542442008	46.9033240263058960
dpmaxdo.3 -> maxhr.3	84.0000000000	128.000000	15.053763	58.33333333333	99.99615147714318	81.9021790430464590
maxhr.3 -> baseEF.1	128.000000000	80.0000000	22.939068	10.93750000000	33.8873445987701	12.3471268562275060
baseEF.1 -> maxhr.3	80.0000000000	128.000000	14.336917	17.50000000000	28.9727151393890	20.4322975199196330

FIGURE 3.3 – Extrait du fichier transaction.out correspondant au jeu de données.

Le graphe d'implication est constitué en ne représentant que les implications dont la valeur d'intensité d'implication est supérieure à un seuil choisi par l'utilisateur. Notre approche ajoute le

seuil de confiance. Les règles en rouge dans (Figure 3.3) correspondent à des règles d'implication très forte en termes d'intensité d'implication pour lesquelles la confiance est faible, c'est-à-dire que le nombre d'individus qui les réalisent est relativement faible. C'est ici qu'on peut voir l'intérêt de notre approche, qui donne la possibilité à l'expert premièrement d'avoir l'information sur la confiance de chaque règle et la possibilité de retenir ces règles ou de les ignorer.

3.3.1 Résultats obtenus sans spécifier un seuil de confiance

Dans le graphe implicatif nous avons offert à l'utilisateur la possibilité d'afficher la valeur de confiance de chaque règle. Dans (Figure 3.4), vu le grand nombre de règles, nous avons montré qu'une partie du graphe d'implication. Dans les graphes d'implications qui suivent, la flèche rouge désigne un seuil d'intensité d'implication égale à 99, la flèche bleue un seuil d'intensité d'implication supérieur ou égal à 95 et la flèche verte désigne un seuil d'intensité d'implication de supérieur ou égal à 90.

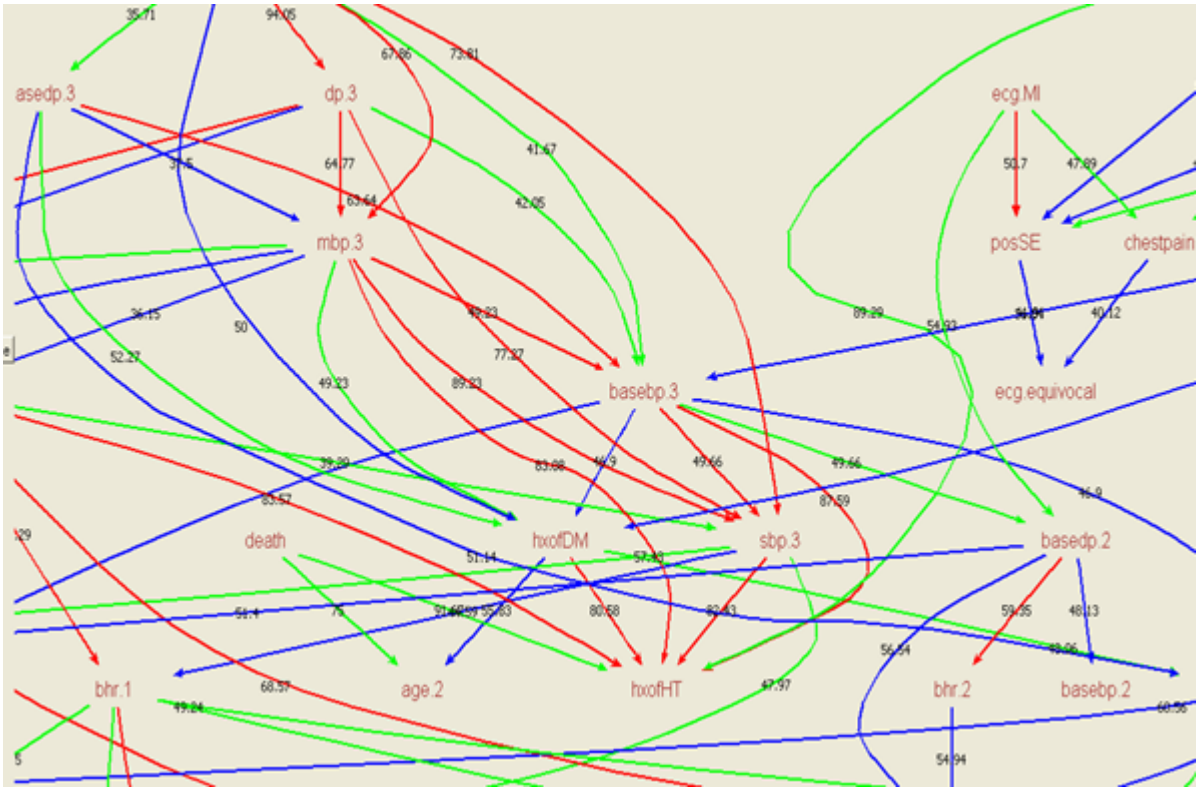

FIGURE 3.4 – Extrait du graphe d'implication avec l'affichage des valeurs de confiance.

En affichant les valeurs de la confiance, l'utilisateur peut voir l'information sur le nombre de participation des individus à chaque règle, mais vu le grand nombre d'implications l'utilisateur aura des difficultés pour distinguer les règles. On a l'impression que tout est lié. Pour cela nous avons rajouté la possibilité de choisir un seuil de confiance qui nous donne la possibilité de voir le niveau d'importance des règles. Dans ce qui suit nous allons présenter le graphe d'implication en affichant juste les implications qui correspondent au seuil de confiance désiré.

3.3.2 Résultats obtenus en utilisant un seuil de confiance égal à 80

Le choix d'un indice de confiance élevé va nous permettre d'extraire les règles d'implications dont le taux de participation est très élevé, cela revient à identifier les facteurs les plus importants qui conduisent à un état de stress. Ci-dessous le graphe d'implication correspondant

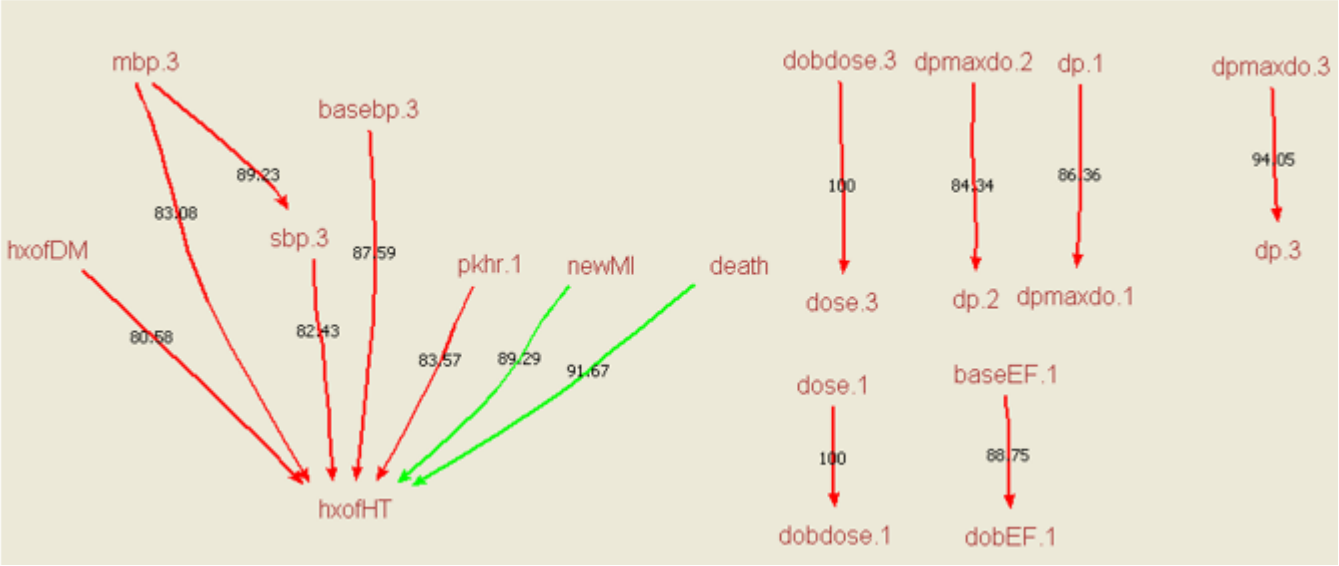

FIGURE 3.5 – Graphe implicatif avec un seuil de confiance égale à 80.

Avec une valeur de confiance 80 nous remarquons que le graphe est plus lisible et que les résultats sont cohérents car les valeurs faibles impliquent les valeurs faibles et les valeurs élevées impliquent les valeurs élevées. Parmi les implications nous rencontrons les suivantes :

- Si la personne est diabétique alors elle a un grand risque de se retrouver en état de stress.
- Si la fréquence cardiaque de base est faible alors la personne a tendance à être en état de stress.
- Si la pression sanguine maximale et la tension de base sont très élevées alors la personne risque de se retrouver en état de stress.
- Si la personne subit une crise cardiaque c'est qu'elle a tendance à se retrouver en un état de stress.

À partir de ces implications nous constatons que les facteurs qui conduisent le plus à un état de stress sont le fait d'avoir :

- Le diabète
- Une fréquence cardiaque faible
- Une pression sanguine maximale très élevée
- Une tension de base élevée
- Une nouvelle crise cardiaque ;

Car les valeurs de confiance sont très élevées.

3.3.3 Résultats obtenus en utilisant un seuil de confiance égale à 70

La figure ci-dessous montre le graphe d'implication correspondant aux résultats obtenus avec un seuil de confiance égale à 70

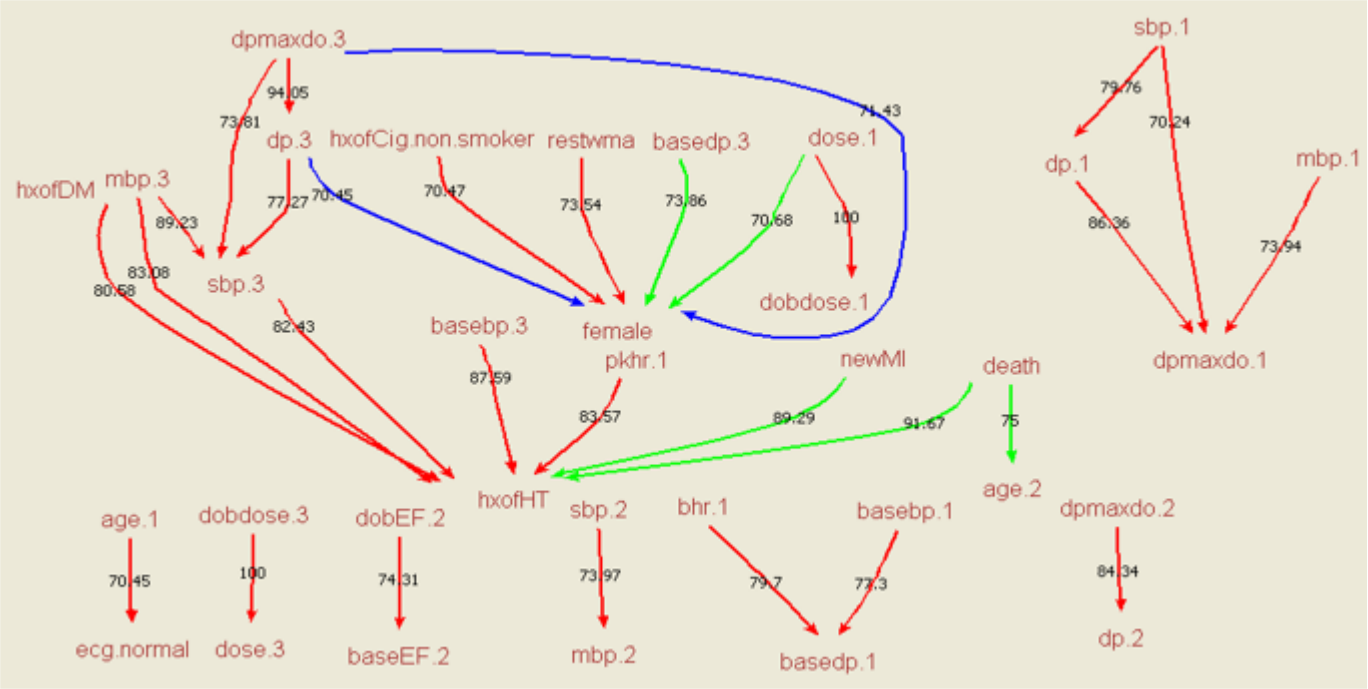

FIGURE 3.6 – Graphe implicatif avec un seuil de confiance égale à 70.

Avec un seuil de confiance égale à 70, nous remarquons qu'on a, en plus des règles dans le graphe précédent, d'autres règles mais avec un seuil de confiance moins élevé, il s'agit des implications suivantes :

- Les médecins ont utilisé une dose très élevée du double de dobutamine pour 71% des femmes.
- Dans le jeu de données utilisé, plus de 70 % des personnes non-fumeurs sont des femmes.
- 73% des personnes qui ont l'anomalie de mouvement de la paroi au repos sont des femmes.
- Les personnes très jeunes ont un diagnostic normal.

3.3.4 Résultats obtenus avec un seuil de confiance égal à 65

La figure ci-dessous montre le graphe d'implication correspondant aux résultats obtenus avec un seuil de confiance égale à 65.

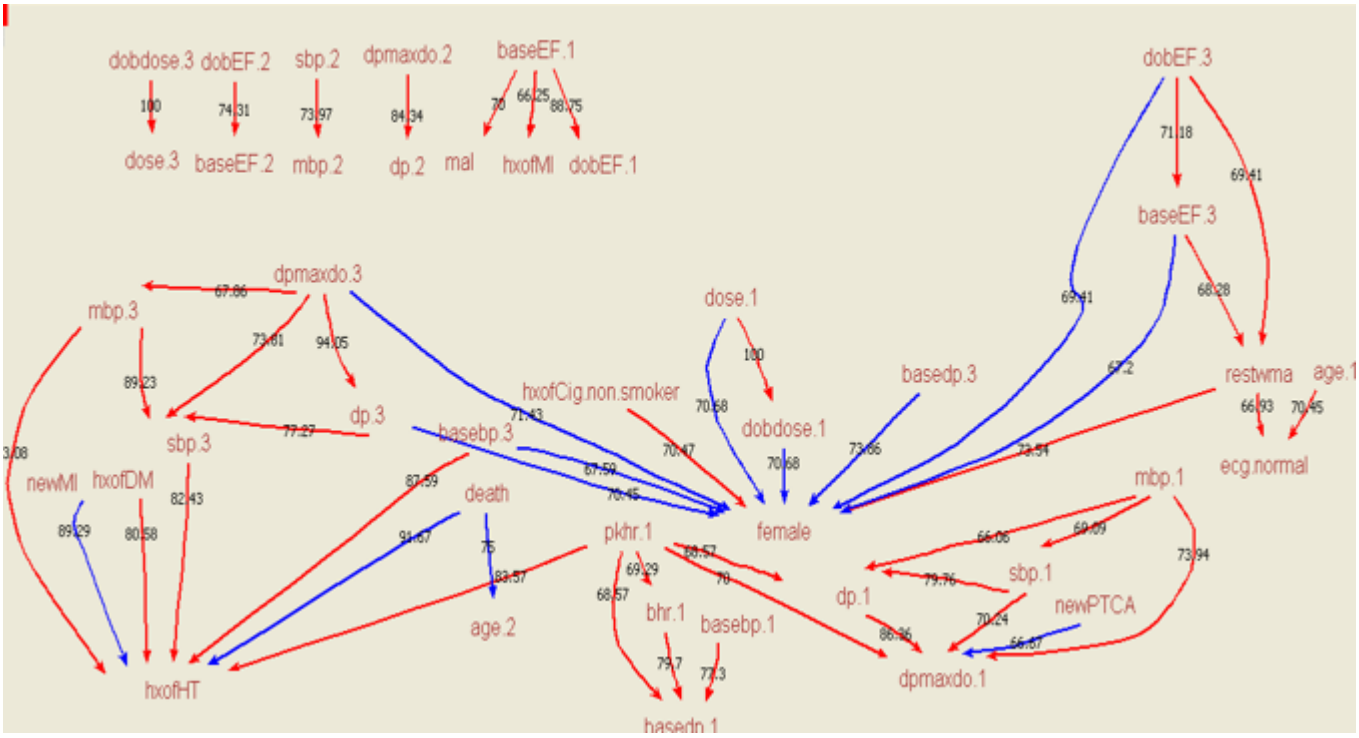

FIGURE 3.7 – Graphe implicatif avec un seuil de confiance égale à 65.

En diminuant le seuil de confiance à 65, plus d'implications apparaissent parmi lesquelles on voit que :

— 66% des personnes qui ont l'anomalie de mouvement de la paroi au repos ont tendance à avoir un diagnostic ecg normale.

— 68 % des personnes qui ont la fraction d'injection cardiaque initiale très élevée et 69% qui ont la fraction d'injection cardiaque élevée sur dobutamine ont tendance à avoir l'anomalie de mouvement de la paroi au repos.

3.4 Conclusion

Dans cette contribution nous avons étudié le couplage de l'indice d'implication avec la confiance afin d'améliorer et de faciliter la recherche de règle avec la méthode d'analyse statistique implicative. Notre but est d'aider les experts et utilisateurs à utiliser le graphe implicatif. Pour cela, nous avons rajouté la valeur de la confiance pour chaque règle dans le graphe, ce qui permet de distinguer le niveau d'importance de chaque règle. Nous avons aussi rajouté un seuil de confiance afin de rendre le graphe plus lisible. Après avoir défini l'approche et l'avoir mise en œuvre dans le logiciel de statistique RCHIC, nous avons appliqué notre approche sur différents ensembles de données. Dans ce papier nous avons choisi les données issues des échocardiographies de stress. Les résultats obtenus nous ont permis d'avoir des graphes plus lisibles et d'identifier et de classer les facteurs les plus importants qui conduisent à un état de stress.

Chapitre 4

Classification en utilisant les règles d'implication de l'ASI

Sommaire

4.1 Introduction 58
4.2 Pourquoi l'ASI pour faire de la classification 58
4.3 Description des ensembles de données 60
4.4 Critères d'évaluation 61
4.5 Approche 1 : Classification avec partitionnement de données 63
4.5.1 Description de l'approche 63
4.5.2 Exemple d'application de l'approche 66
4.5.3 Expérimentation 70
4.5.4 Evaluation 71
4.6 Approche 2 : Classification sans partitionnement de données 74
4.6.1 Description de l'approche 74
4.6.2 Implémentation de l'approche 76
4.6.3 Exemple d'application 77
4.6.4 Résultats et discussions 78
4.7 Conclusion 82

4.1 Introduction

Les méthodes de classifications peuvent être divisées en deux catégories : simples ou complexe. Les classificateurs simples sont plus interprétables mais généralement moins précis (par exemple, CART, C4.5, J48). Les classificateurs complexes sont moins interprétables mais plus précis (par exemple, les réseaux de neurones, SVM). Dans ce compromis entre interprétabilité et précision, nous proposons deux nouvelles méthodes de classifications à la fois faciles à interpréter et conduisant à une précision pertinente. Ces deux méthodes de classification, sont basées sur le concept d'intensité d'implication de l'analyse statistique implicative (SIA). L'objectif de ces deux approches est de définir un classificateur binaire précis et facile à interpréter, et de montrer qu'il est aussi compétitif que les autres classificateurs plus complexes et largement utilisés. Pour cela, nous avons utilisé une mesure de qualité issue de l'analyse statistique implicative (voir section 1.2.4.4 page 18).

Dans la suite de ce chapitre nous allons, justifier notre choix d'utilisation de l'ASI pour faire de la classification, donner une brève description des jeux de données en libre accès utilisés pour le test. Puis nous allons présenter les deux méthodes de classifications et la logique de chaque classificateur sera exposé.

4.2 Pourquoi l'ASI pour faire de la classification

Nous montrons l'intérêt de l'utilisation de l'ASI spécialement l'indice d'implifiance (qui combine l'intensité d'implication et la confiance), pour faire de la classification. Nous allons rappeler brièvement les définitions des mesures de qualité les plus utilisés dans les algorithmes basés sur les règles d'association, tout en exposant leurs points fort et leurs faiblesses. Nous allons montrer le rôle important de l'intensité d'implication associé avec la confiance pour surmonter les différents problèmes de ces indices.

Les règles d'association sont dotées de plusieurs mesures de qualité. Les plus utilisées sont la confiance, le support et le lift. Elles peuvent être facilement affichés lorsque les règles sont extraites par l'algorithme Apriori [GSGS08], mis en œuvre dans le package arules du logiciel R [HCHB11]. Il existe plusieurs autres mesures de qualité, chacune permettant d'extraire une information légèrement différente de la règle considérée [VLL08] [LVL07] [LLV05]. Pour obtenir de bons taux de prédiction, la confiance est la plus importante, car elle mesure la probabilité que le côté droit (conclusion ou rhs) de la règle soit observé pour un individu détenant le côté gauche (prémisse ou lhs) de la règle. En utilisant le support, le calcul prend beaucoup de temps pour les grand ensembles de données, et l'algorithme Apriori doit négliger les règles affectant une faible proportion de l'échantillon, afin d'accélérer le temps de calcul. Le lift d'une règle peut être interprété comme l'effet de la lhs sur la rhs : On peut avoir des règles avec un niveau de confiance de 0,80 (ce qui donne une bonne capacité de prédiction), mais un lift de seulement 0,90, signifie que la lhs réduit les chances initiales de la rhs. Pour avoir une meilleure prédiction il ne suffit

	$\overline{b}$	b	Total
$\overline{a}$	55	5	60
a	20	20	40
Total	75	25	100

TABLEAU 4.1 – Le tableau de fréquence conjointe des variables binaires a et b.

pas d'avoir des mesures qui prennent en considération le nombre d'occurrence de l'ensemble de départ,(comme la confiance), mais aussi celles qui prennent en compte le nombre d'occurrence de l'ensemble d'arrivée.

Toutes ces mesures de qualité sont descriptives [1], donc sujettes à une variabilité naturelle dans le cadre du plan d'échantillonnage. Afin de mesurer l'effet des lhs sur les rhs, on peut utiliser l'inférence statistique [2] sur la valeur de lift d'une règle dans la population entière. Dans ce but, l'intensité d'implication est utilisée. L'intensité d'implication de la règle $a \rightarrow b$ tel qu'on l'a déjà défini est la probabilité que la relation entre a et b produise plus de contre-exemples (dans le cas d'absence de lien entre a et b) que ceux observés dans l'échantillon. De manière formelle, $P(N_{a\bar{b}} > n_{a\bar{b}})$, où $N_{a\bar{b}}$ est une variable aléatoire représentant le nombre de contre-exemples observés dans un échantillon aléatoire, et $n_{a\bar{b}}$ représentant le nombre de contre-exemples observés dans l'échantillon. Selon le plan d'échantillonnage aléatoire, la distribution de $N_{a\bar{b}}$ peut être binomiale ou poissonienne, et aussi une approximation gaussienne est possible pour les échantillons de grande taille [GSGS08]. D'autres approches probabilistes pour définir l'intérêt des règles sont présentées dans [TF+19].

Exemple 1. Considérant les variables binaires a et b ayant la distribution conjointe donnée dans le tableau 4.1, l'intensité d'implication de $A \rightarrow B$ est $P(N_{a\bar{b}} > 20)$. Nous pouvons considérer A comme le résultat d'essais de Bernoulli de paramètre p = 40/100, et de même pour $\overline{B}$, avec le paramètre p = 75/100. En vertu de l'indépendance statistique, nous pouvons voir les observations des contre-exemples (A = 1, B = 0) comme des essais de Bernoulli de paramètre p = 0.4. 0.75= 0.3. En appliquant la loi de distribution binomiale avec le paramètre de Bernoulli de 0.3, la probabilité $P(N_{a\bar{b}} > 20)$ sera égale à $0,9835$ Dans l'hypothèse d'une indépendance statistique entre A et B, le nombre observé de contre-exemples dépasse celui observé dans l'échantillon avec la probabilité $P(N_{a\bar{b}} > 20) = 0,9835$. Bien que la confiance de la règle est faible, $20/40 = 0,5$, le lift est de $0,5/0,25 = 2$, ce qui montre le fait qu'un ensemble d'éléments contenant $A(A = 1)$ double les chances que cet ensemble contient B $(B = 1)$. L'implication explique cet effet significatif de A sur B. Le tableau de fréquence conjointe des variables binaires a et b (Tableau 4.1) illustre les concepts de confiance, de lift et d'intensité d'implication.

Remarque. L'intensité d'implication coïncide avec la (complément à 1 de la) p-value du test

1. On appelle mesure descriptive une mesure qui ne change pas en cas de dilatation des données.
2. L'inférence statistique consiste à induire les caractéristiques inconnues d'une population à partir d'un échantillon issu de cette population.

d'hypothèse statistique H_0 : quand lift ≤ 1 vs H_1 : quand lift > 1, lorsque l'on utilise la statistique de test $N_{a\bar{b}}$, et la procédure rejette H_0 si $N_{a\bar{b}} \geq c$ pour un certain c.

En gardant cela à l'esprit, nous construisons deux classificateurs basés sur des règles fortes, montrant une grande confiance, mais aussi des règles significatives. De plus, l'objectif est de choisir des règles dont la lhs montre un effet positif significatif sur la rhs (c'est-à-dire, quand l'intensité d'implication est forte). Ensuite, la notion d'implifiance en tant que mesure de qualité utile ayant une grande valeur de confiance.

4.3 Description des ensembles de données

Pour analyser les performances de l'ASI dans la classification, plusieurs jeux de données parmi les plus utilisés ont été testés par nos approches. Ces ensembles de données contiennent des éléments (individus, objets,...) qui nécessitent des traitements où chaque traitement est représenté par une variable (par exemple la variable V3 de l'ensemble de données WBC représente un traitement qui consiste à contrôler la taille de la cellule). Les résultats obtenus par ces traitements pour chaque élément sont stockés dans des ensembles de données. La première approche a été testée avec trois jeux de données sur le cancer de sein (Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC), Wisconsin Prognosis Breast Cancer (WPBC)). Le cancer du sein est la deuxième cause du décès chez les femmes après le cancer des poumons [Gro14]. La deuxième approche a été également testée par rapport à ces trois ensembles déjà cité, plus trois autres. Haberman's survival data set (Haberman), EEG Eye State Data Set (IDs-mapping) et Cervical Cancer Behavior Risk Data Set (SOBAR) [AN07]. Une brève description de ces ensembles de données est présentée ci-dessous.

Le jeu de données WBC, permet de désigner les tumeurs maligne et bénigne (2 pour bénigne, 4 pour maligne), composé de 10 attributs : numéro d'identification, épaisseur des cellules, uniformité de la taille de la cellule, uniformité de la forme cellulaire, adhérence marginale, taille de cellule épithéliale unique, noyau nu, chromatine du noyau, noyaux normaux, mitoses et le diagnostic (maligne, bénigne). Il comporte 683 entrées avec 444 distributions de classe négative et 239 positive.

Le diagnostic (M = maligne, B = benigne) avec le jeu de données WDBC est relié à dix caractéristiques pour chaque noyau cellulaire, ces caractéristiques sont : rayon, texture, périmètre, zone, lissage, compacité, concavité, points concaves, symétrie et dimension fractale. Toutes les caractéristiques sont modélisées numériquement, de sorte qu'une plus grande valeur indiquera généralement une plus grande probabilité d'atteinte du cancer. En résumé le jeu de données comporte 569 enregistrements avec 30 attributs, 212 distributions de classe positive et 357 négatives.

Dans le jeu de données WPBC les sortie (R = récurrent, N = non récurrent) sont reliées aux temps (temps de récurrence, pour récurrent, et temps sans maladie pour non-récurrent). Avec les mêmes attributs que le jeu de données WDBC plus deux autres, la taille des tumeurs et les

stades des ganglions lymphatiques. L'ensemble de données comporte 194 enregistrements avec 46 distributions de classe positive et 148 de classe négative.

Le jeu de données Haberman reflète une étude sur la survie des patientes ayant subi une opération chirurgicale pour un cancer de sein. Il contient 306 instances et 3 attributs (l'âge de la patiente au moment de l'opération, l'année où la patiente a été opérée et le nombre de ganglions auxiliaires positifs détectés) ainsi que la classe (1 = la patiente a survécu 5 ans ou plus, et 2 = la patiente est décédée dans les 5 ans). Il comporte 81 enregistrements positifs et 225 négatifs.

La classification de l'état des yeux EEG (IDs-mapping), est importante et utile pour détecter l'état de cognition des humains. L'ensemble de données comprend 14 mesures EEG continues. L'état des yeux a été détecté par une caméra pendant la mesure de l'EEG et ajouté manuellement au fichier après analyse des images vidéo. Un '1' indique que les yeux sont fermés et un '0' qu'ils sont ouverts, avec un total de 6723 enregistrements positifs et 8257 négatifs.

L'ensemble de données sur le risque de cancer du col de l'utérus SOBAR contient 18 attributs explicatifs concernant le col de l'utérus ('1'= a un cancer du col de l'utérus, '0' = pas de cancer du col de l'utérus) contenant 72 instances, avec 21 enregistrements positifs et 51 négatifs [MW+16].

4.4 Critères d'évaluation

Les deux approches proposées seront comparées à d'autres méthodes de classification largement utilisées dans la littérature (Naïve Bayes, réseaux de neurones à base radiale, arbres de décision J48 et CART) avec les ensembles de données en accès libre provenant du dépôt Machine Learning de l'UCI. Ces ensembles de données ont été utilisés pour tester des classificateurs, comme dans [LHM+98a] [IUY13] [SAZ12] [IHSM15] [IUY13] [O+96]. En outre, l'ensemble de données Haberman a été utilisé pour tester les algorithmes dans [SK13]. L'ensemble de données EEG sur l'état des yeux (IDs-mapping) a été utilisé dans [SMK] pour la classification de l'état des yeux à l'aide de l'algorithme des k-proches voisins et du modèle de réseaux de neurone à perceptron multicouche, et également utilisé dans [TW14], ou une nouvelle approche d'identification de l'état des yeux EEG basée sur l'apprentissage incrémental des attributs est proposé. L'ensemble de données sur le risque comportemental du cancer du col de l'utérus (SOBAR) a également été utilisé dans [HFC20]. La comparaison de nos classificateurs est effectuée par le calcul des matrices de confusion [HPK11] : exactitude, précision, sensitivité et spécificité. Tous les calculs effectués dans ces deux approches ont été effectués à l'aide du logiciel statistique R. Le niveau de pertinence du modèle de classification est calculé avec le nombre de classification correcte et incorrect de chaque valeur possible de variable classifiée dans la matrice de confusion [AB94]. Dans le cadre de notre étude, les entrées dans la matrice de confusion ont les significations suivantes :

— TP est le nombre de vrais positifs (le nombre de prédictions correctes dont l'instance est positive),

— TN est le nombre de vrais négatifs (le nombre de prédictions correctes dont l'instance est

négative),

— FP est le nombre de faux positifs (le nombre de prédictions fausses dont l'instance est positive) et

— FN le nombre de faux négatifs (le nombre de prédictions négatives dont l'instance est négative). Les valeurs de performances utilisées sont données par les formules suivantes [SAZ12] [IHSM15] [IUY13] :

$$Exactitude = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$sensitivite = \frac{TP}{TP + FN}$$

$$Specificite = \frac{TN}{TN + FP}$$

Pour évaluer la capacité de généralisation d'un modèle, il est nécessaire de le tester sur de nouvelles données. Par ailleurs, moins on a de données, plus l'apprentissage est limité. Pour cela nous avons utilisé la validation croisée "cross-validation" dans la deuxième approche, qui nous a permis d'utiliser l'intégralité du jeu de données pour l'entraînement et pour la validation.

La "cross-validation" est une méthode statistique qui permet d'évaluer la capacité de généralisation d'un modèle. Elle est utilisée pour tester l'efficacité d'un modèle de Machine Learning. Généralement lorsqu'on parle de cross-validation (CV), l'on réfère à sa variante la plus populaire qu'est le k-fold cross-validation. Cette méthode est une procédure de « rééchantillonnage » permettant d'évaluer un modèle même avec des données limitées. Son principe consiste à découper le jeu de données en k parties à peu près égales. Chacune des k parties est utilisée comme jeu de test. Le reste (autrement dit, l'union des k-1 autres parties) est utilisé pour l'apprentissage. On utilise couramment la Cross-Validation pour comparer différents modèles et sélectionner le plus approprié pour un problème spécifique. Elle est à la fois simple à comprendre, simple à implémenter que les autres méthodes. Pour cela nous l'avons utilisé pour mieux évaluer notre approche.

4.5 Approche 1 : Classification avec partitionnement de données

Dans cette approche, notre préoccupation majeure est d'avoir des règles qui ont du sens. Ces règles seront présentées aux experts afin de prédire les classes des différents éléments d'un ensemble de données de manière facile et avec une erreur de prédiction très faible. Dans la suite, nous exposons la logique de ce classificateur, nous donnons un exemple d'application détaillé, nous expérimentons le classificateur avec les différents jeux de données, nous évaluons le classificateur en le comparant à d'autres outils du Data Mining, finalement nous concluons avec une conclusion.

4.5.1 Description de l'approche

Dans cette section nous décrivons avec détail toutes les étapes nécessaire pour la réalisation du premier classificateur basé sur les règles d'implication de l'ASI. L'approche consiste à :

- Préparer et traiter les données d'entrées.
- Extraire les règles d'implication à partir des données en utilisant l'ASI.
- Filtrer les règles extraites et sélectionner les variables (règles) significatives.
- Définir plusieurs variables (règles) pour faire les prédictions.
- Former les combinaisons de variables (règles).
- Choisir la combinaison qui minimise l'erreur de prédiction.
- Simplifier les règles obtenues.

Préparation et traitement des données d'entrées

Les ensembles de données traités par notre approche contiennent des éléments (individus, objets ...) ayant subi des traitements. Les résultats obtenus par ces traitements sont représentés par des variables. Chaque variable dans le jeu de données est partitionnée en un nombre fixe d'intervalles. Le nombre d'intervalles est sélectionné par l'utilisateur, ensuite l'algorithme des nuées dynamiques Diday (1971)[Did71] constitue automatiquement les intervalles qui ont des limites distinctes. Généralement nous choisissons le partitionnement par défaut, ce qui veut dire que chaque intervalle est partitionné en trois sous intervalles. Les variables avec des valeurs numériques sont partitionnées pour obtenir des variables binaires. Elles sont divisées en un petit nombre d'intervalles (par exemple, la variable V_i peut être divisée en ni intervalles $I_1, I_2, ..., I_{ni}$). Pour chaque $j = 1, 2, ..., ni$, on peut définir la variable binaire $V_{i,j}$ (c'est-à-dire n'atteignant la valeur 1 que pour les données dont la variable V_i appartient à l'intervalle I_j, et 0 ailleurs) (Voir Figure 4.1). Le nombre d'intervalles pour les variables numériques est le premier paramètre de notre classificateur.

Extraction des règles d'implication à partir des données d'entrées

Avec l'ASI, les données sont analysées afin de trouver des relations entre l'appartenance d'un élément à un échantillon donné d'une variable et son appartenance à une classe donnée. Ces relations sont formalisées sous forme de règles d'implications. Par exemple, l'ASI peut construire des implications de la forme suivante : si un élément appartient à une partition donnée d'une variable donnée, cela implique généralement qu'il appartient à une autre partition d'une autre variable. Par exemple avec l'ASI, la règle ($V_{1.1} \Rightarrow V_{5.2}$) signifie que si un individu appartient à la partition 1 de la variable V_1 alors il appartient généralement aussi à la partition 2 de la variable V_5. L'ASI utilise l'algorithme Apriori pour extraire ces régles et les enregistre dans un fichier appelé "transaction.out" (Voir Figure 4.2).

Filtrage des règles extraites et sélection des variables (règles) significatives.

Notre objectif est de pouvoir prédire la classe de nouveaux éléments avec les règles sélectionnées précédemment. Les questions qui se posent sont : Est-ce que toutes les règles obtenues sont intéressantes ? Quelle est la règle à utiliser pour faire la prédiction ? Est-ce que la prédiction avec une seule règle donne la meilleure prédiction ? A partir des règles précédemment obtenus, nous nous intéressons à :

- Toutes les règles possibles pour lesquelles la conclusion représente un nom d'une classe (toutes les règles avec lesquelles nous pouvons faire la prédiction).
- Parmi ces règles, nous sélectionnons uniquement les règles les plus importantes. Pour cela nous devons définir un seuil. Nous utilisons l'implifiance pour détecter l'importance de chaque règle. La notion d'implifiance est présentée dans le deuxième chapitre comme la mesure de qualité qui tient compte à la fois de la confiance (la meilleure pour la prédiction) et d'un paramètre statistiquement significatif de la lhs sur la rhs (intensité d'implication). Chaque règle de classification de longueur 2 (i.e., dont la lhs est une variable d'entrée binaire et dont la rhs est soit $Y0$ soit $Y1$) peut être extraite. Le seuil d'implifiance est le deuxième paramètre du modèle, et il filtrera les règles afin de ne garder que les plus significatives. Une fois les règles significatives détectées, les variables concernées sont les variables significatives pour le processus de classification. Pour retenir une variable, toutes les règles contenant en prémisse une partition de cette variable doivent dépasser le seuil d'implifiance choisi.

Plusieurs variables (règles) pour faire les prédictions

Dans notre approche, au lieu d'utiliser une seule règle (ou une seule variable car chaque règle contient une variable) pour faire les prédictions, nous utilisons une combinaison de plus d'une règle (variable) car, il peut y avoir plusieurs facteurs qui influence l'état de chaque élément. Par exemple, si nous avons deux classes (C et D) et nous avons les règles suivantes : $V_{1.1} \Rightarrow C$, $V_{1.2} \Rightarrow D$, $V_{2.1} \Rightarrow C$, $V_{2.2} \Rightarrow D$, $V_{5.2} \Rightarrow C$, $V_{5.2} \Rightarrow C$, $V_{6.1} \Rightarrow C$, $V_{6.2} \Rightarrow D$, $V_{8.1} \Rightarrow D$, $V_{8.2} \Rightarrow D$. Alors nous

remarquons que les classes C et D sont donnés par plus d'une variable (facteur) pour le même élément. Dans cet exemple, un élément appartient à la classe C ou D selon son appartenance à la classe une ou deux de chaque variable.

Formation des combinaisons de variables (règles)

Plus le nombre de variables utilisées pour définir le processus de classification est élevé, plus l'erreur de prédiction est faible. A l'inverse, la méthode de classification résultante sera moins compréhensible pour le praticien. Afin d'obtenir un compromis, un troisième paramètre doit être choisi, un nombre impair (nvote) pour regrouper les variables significatives.

par défaut l'approche utilise une combinaison de trois variables parce qu'avec trois variables nous obtenons un nombre faible de règles à présenter aux experts pour prédire la classe des éléments. Nous sélectionnons toutes les combinaisons de trois variables parmi toutes les variables choisis (par exemple si nous avons les variables $(V_1, V_2, V_5, V_6 \, et \, V_8)$ nous obtenons les combinaisons suivantes $(V_1, V_2, V_5), (V_1, V_5, V_8)$ ..., etc.)(Voir Figure 4.4). Si nous avons n variables choisies, le nombre de k-combinaisons (k égale à 3 dans notre approche) est obtenu avec cette formule :

$$\frac{n\,!}{3\,!(n-3)\,!} \quad si \quad k \leq n.$$

Choix de la combinaison qui minimise l'erreur de prédiction

Pour calculer l'erreur de prédiction de chaque combinaison, nous calculons l'erreur de prédiction en utilisant chaque règle dans la combinaison. Pour cela nous allons calculer :

- La prédiction de chaque individu par rapport à chaque règle dans la combinaison (Voir Figure 4.5)
- La prédiction de chaque individu par rapport à chaque combinaison (en prenant la prédiction majoritaire : vote majoritaire)
- L'erreur de prédiction par rapport à chaque combinaison de variable
- La combinaison avec l'erreur de prédiction la plus petite. (Voir Figure 4.7)

Par exemple, si nous avons la combinaison (V_1, V_2, V_8) nous calculons l'erreur de prédiction en utilisant les règles correspondantes : $(V_{1.1} \Rightarrow C, V_{1.2} \Rightarrow D, V_{2.1} \Rightarrow C, V_{2.2} \Rightarrow D, V_{8.1} \Rightarrow D, V_{8.2} \Rightarrow D)$. Pour prédire la classe d'un élément en utilisant une seule variable, nous prenons la règle qui a en prémisse la partition de la variable, la conclusion représente la prédiction. Par exemple, si un individu appartient à la partition 1 de la variable $V_1, V_2 \, et \, V_8$. La prédiction en utilisant chaque variable est déduite à partir de la conclusion des règles $V_{1.1} \Rightarrow C, V_{2.1} \Rightarrow C$ and $V_{8.1} \Rightarrow D$. Nous prenons la conclusion la plus répétée.Dans notre cas c'est la classe C.

Simplifier les règles utilisées pour faire les prédictions.

Les règles obtenues peuvent être simplifiés pour avoir moins de combinaisons avec moins de conjonctions (Figure 4.7, Figure 4.8).

En résumé les étapes de l'approche sont :

— Transformer chaque variable d'entrée V_i en un ensemble de variables binaires $V_{i.1}, V_{i.2}, ..., V_{i.ni}$ (on a ici un paramètre, le nombre d'intervalles dans lequel chaque variable numérique est partitionnée).

— La variable de classification Y est doublée en variables binaires $Y0$ et $Y1$.

— Exploiter toutes les règles de classification de longueur 2 (dont le côté gauche est une variable d'entrée binaire et le côté droit est soit $Y0$, soit $Y1$). Ce choix nous permet d'extraire toutes les règles sans ignorer celle dont le support minimum est faible, même pour les grands ensembles de données, dans un temps raisonnable.

— Choisir un seuil $i0$ (il s'agit d'un nouveau paramètre dans $[0, 1]$) pour l'implifiance, et filtrez les règles qui dépassent cette implifiance. Toute variable originale V_i telle que toutes ses variables binaires liées ($V_{i.1}, V_{i.2}, ..., V_{i.ni}$) sont les prémices d'une règle significative peut être appelée variable significative. Les variables ayant un faible effet sur la variable de classification Y sont rejetées.

— Choisir le nombre de règle à utiliser pour faire de la prédiction noté « *nvote* » par défaut il est égal à 3.

— Former toutes les combinaisons de « *nvote* » variables possible.

— Calculer l'erreur de prédiction en utilisant chaque combinaison, Pour cela nous allons calculer la prédiction de chaque individu par rapport à chaque règle dans la combinaison. Nous calculons également la prédiction de chaque individu par rapport à chaque combinaison (en prenant la prédiction majoritaire). Ensuite nous calculons l'erreur de prédiction par rapport à chaque combinaison de variable.

— Prendre la combinaison avec l'erreur de prédiction la plus petite

— Simplifier les règles obtenues pour avoir moins de combinaisons avec moins de conjonctions.

Avec notre approche, nous obtenons un nombre faible de règles logiques qui sont compréhensibles par l'être humain. Ces règles seront présentées aux experts pour faciliter la classification des ensembles de données.

4.5.2 Exemple d'application de l'approche

Dans cet exemple, nous allons appliquer notre approche au domaine de la médecine. Nous utilisons l'ensemble de données WBC (voir section 4.3 page 60). Dans ce jeu de données, chaque variable est partitionnée en échantillon. La figure ci-dessous montre un extrait du jeu de données WBC partitionnée en deux échantillons.

	V1.2	V2.1	V2.2	V3.1	V3.2	V4.1	V4.2	...	V9.1	V9.2	malignant	benign
1	1	1	0	1	0	1	0		1	0	0	1
2	1	1	0	1	0	0	1		1	0	0	1
3	0	1	0	1	0	1	0		1	0	0	1
4	1	0	1	0	1	1	0		1	0	0	1
5	0	1	0	1	0	1	0		1	0	0	1
6	1	0	1	0	1	0	1		1	0	1	0
7	0	1	0	1	0	1	0		1	0	0	1
8	0	1	0	1	0	1	0		1	0	0	1
9	0	1	0	1	0	1	0		0	1	0	1
10	0	1	0	1	0	1	0		1	0	0	1
11	0	1	0	1	0	1	0		1	0	0	1
12	0	1	0	1	0	1	0		1	0	0	1
13	1	1	0	1	0	1	0		1	0	1	0

FIGURE 4.1 – Extrait du jeu de données WBC partitionné en deux échantillons.

Les règles d'implication sont générées en utilisant l'algorithme Apriori, elles sont représentées dans le fichier transaction.out, représenté ci-dessous.

row.names	occurrence.hyp.	occurrence.con.	support.rule.	confidence	classical.index		classical.simi
V3.2 -> malignant	183	239	26.793558	96.1748634	1.000000e+02	9.863704e+01	1.000000e+02
malignant -> V2.2	239	175	34.992679	71.9665272	1.000000e+02	8.890503e+01	1.000000e+02
V2.2 -> malignant	175	239	25.622255	98.2857143	1.000000e+02	9.940003e+01	1.000000e+02
malignant -> V6.2	239	174	34.992679	70.2928870	1.000000e+02	8.818951e+01	1.000000e+02
V6.2 -> malignant	174	239	25.475842	96.5517241	1.000000e+02	9.878996e+01	1.000000e+02
malignant -> V7.2	239	173	34.992679	67.7824268	1.000000e+02	8.709811e+01	1.000000e+02
V7.2 -> malignant	173	239	25.329429	93.6416185	1.000000e+02	9.775601e+01	1.000000e+02
malignant -> V8.2	239	155	34.992679	60.2510460	1.000000e+02	8.384065e+01	1.000000e+02
V8.2 -> malignant	155	239	22.693997	92.9032258	1.000000e+02	9.756273e+01	1.000000e+02
malignant -> V5.2	239	144	34.992679	55.6485356	1.000000e+02	8.176891e+01	1.000000e+02
V5.2 -> malignant	144	239	21.083455	92.3611111	1.000000e+02	9.742009e+01	1.000000e+02
malignant -> V4.2	239	141	34.992679	55.6485356	1.000000e+02	8.179660e+01	1.000000e+02
V4.2 -> malignant	141	239	20.644217	94.3262411	1.000000e+02	9.810339e+01	1.000000e+02

FIGURE 4.2 – Extrait du fichier transaction.out.

Nous utilisons un seuil d'implifiance égal à $0,8$. En premier lieu, nous sélectionnons les règles dont la valeur d'implifiance est supérieure à $0,8$. Pour retenir une variable, toutes les règles contenant en prémisse une partition de cette variable doivent dépasser le seuil d'implifiance. La liste des variables correspondantes à ces règles est la suivante : $(V_1, V_2, V_3, V_4, V_5, V_6, V_7, V_8)$. La variable V_9 n'est pas sélectionnée car la valeur d'implifiance de la règle $V9.1 \Rightarrow$ bénigne est inférieur au seuil sélectionné.

	row.names	occurrence.hyp.	occurrence.con.	support.rule.	confidence	classical.index	
15	V9.1 -> benign	643	444	94.143485	68.5847589	9.374259e+01	5.351122e+01
21	V9.1 -> malignant	643	239	94.143485	31.4152411	1.302704e+01	2.796182e+00
24	V9.2 -> malignant	40	239	5.856515	92.5000000	1.000000e+02	9.790375e+01
270	V9.2 -> benign	40	444	5.856515	7.5000000	8.748555e-06	4.389752e-06

FIGURE 4.3 – Règles obtenus avec la variable V_9 en prémisse.

Afin de sélectionner les combinaisons de règles à appliquer, nous calculons toutes les combinaisons possibles de trois variables parmi les 8 variables retenues. Un extrait de la liste des combinaisons de trois variables parmi 8 est donné dans la figure suivante. Chaque colonne représente une combinaison de variable.

	1	2	3	4	5	6	7	8	.9	10	...	51	52	53	'54	55	56
1	V4	V4	V4	V4	V4	V4	V4	V4	V4	V4		V7	V7	V6	V6	V6	V2
2	V5	V5	V5	V5	V5	V5	V8	V8	V8	V8		V2	V3	V2	V2	V3	V3
3	V8	V7	V6	V2	V3	V1	V7	V6	V2	V3		V1	V1	V3	V1	V1	V1

FIGURE 4.4 – Exemple de liste de combinaison de variables.

Avant d'obtenir l'erreur de prédiction de chaque combinaison de variable, nous calculons en premier lieu la prédiction en utilisant chaque variable. La prédiction en utilisant chaque variable est donnée dans la Figure 4.5

	V4	V5	V8	V7	V6	V2	V3	V1	ETAT
1	benign	benign	benign	benign	benign	benign	benign	malignant	benign
2	malignant	malignant	benign	benign	malignant	benign	benign	malignant	benign
3	benign	benign	benign	benign	benign	benign	benign	benign	benign
4	benign	benign	malignant	benign	benign	malignant	malignant	malignant	benign
5	benign	benign	benign	benign	benign	benign	benign	benign	benign
6	malignant	malignant	malignant	malignant	malignant	malignant	malignant	malignant	malignant
7	benign	benign	benign	benign	malignant	benign	benign	benign	benign
8	benign	benign	benign	benign	benign	benign	benign	benign	benign
9	benign	benign	benign	benign	benign	benign	benign	benign	benign
10	benign	benign	benign	benign	benign	benign	benign	benign	benign
11	benign	benign	benign	benign	benign	benign	benign	benign	benign
12	benign	benign	benign	benign	benign	benign	benign	benign	benign
13	benign	benign	benign	benign	benign	benign	benign	malignant	malignant

FIGURE 4.5 – La prédiction en utilisant chaque variable pour chaque individu.

Par exemple la prédiction de l'individu 1 avec la variable V_2 est benign car l'individu 1 appartient au premier échantillon de la variable $V2$ ($V_{2.1} = 1$ dans la Figure 4.1 et nous avons la règle $V_{2.1} \Rightarrow$ benign dans la liste des règles obtenus.

V4.1 ⇒benign; – V1.2 ⇒malignant;
V5.1 ⇒benign; – V3.2 ⇒malignant;
V8.1 ⇒benign; – V2.2 ⇒malignant;
V7.1 ⇒benign; – V6.2 ⇒malignant;
V6.1 ⇒benign; – V7.2 ⇒malignant;
V2.1 ⇒benign; – V8.2 ⇒malignant;
V3.1 ⇒benign; – V5.2 ⇒malignant;
V1.1 ⇒benign; – V4.2 ⇒malignant.

FIGURE 4.6 – Liste de règles obtenues.

Par la suite, nous pouvons calculer la prédiction en utilisant une combinaison de variable, qui représente la prédiction la plus répétée de toutes les prédictions obtenues avec chaque variable dans la combinaison. Nous calculons l'erreur de prédiction de chaque combinaison, à partir des prédictions de tous les individus par rapport à chaque combinaison. Par exemple la combinaison de variable qui a la plus petite erreur de prédiction est : (V_6, V_2, V_1). Pour chaque combinaison de variable, nous avons plusieurs combinaisons de règles. Pour la combinaison de variable en haut, les combinaisons de règles correspondantes sont données dans la Figure 4.7. La prédiction avec la combinaison est obtenue en prenant la prédiction majoritaire des règles formant la combinaison.

	Regle1	Regle2	Regle3	Prediction
1	V6.1 benign	V2.1 benign	V1.1 benign	benign
2	V6.1 benign	V2.1 benign	V1.2 malignant	benign
3	V6.1 benign	V2.2 malignant	V1.1 benign	benign
4	V6.1 benign	V2.2 malignant	V1.2 malignant	malignant
5	V6.2 malignant	V2.1 benign	V1.1 benign	benign
6	V6.2 malignant	V2.1 benign	V1.2 malignant	malignant
7	V6.2 malignant	V2.2 malignant	V1.1 benign	malignant
8	V6.2 malignant	V2.2 malignant	V1.2 malignant	malignant

FIGURE 4.7 – Les combinaisons de règles correspondantes à la combinaison (V_6, V_2, V_1).

Nous pouvons simplifier ces combinaisons pour avoir moins de combinaisons avec moins de conjonctions. Par exemple, dans les deux premières combinaisons, quelle que soit la valeur de la variable V_1, nous avons toujours $V_{6.1} \Rightarrow$ benign et $V_{2.1} \Rightarrow$ benign, donc au lieu d'avoir deux combinaisons avec trois conjonctions, nous obtenons uniquement une combinaison avec deux conjonctions. La Figure 4.8 représente le résultat de simplification des règles dans la Figure 4.7.

n-partition	n-variable	WBC	WDBC	WPBC
2	3	94.43631	93.49736	74.74747
3	3	96.48609	94.3761	71.21212
3	5	96.92533	95.25483	77,27273
3	7	97.07	94.72759	NA

Tableau 4.2 – Pourcentage d'exactitude pour les ensembles de données WBC, WDBC et WPBC

	Regle1	Regle2	Prediction
1	V2.1 benign	V6.1 benign	benign
2	V2.1 benign	V1.1 benign	benign
3	V6.1 benign	V1.1 benign	benign
4	V6.2 malignant	V1.2 malignant	malignant
5	V2.2 malignant	V1.2 malignant	malignant
6	V2.2 malignant	V6.2 malignant	malignant

Figure 4.8 – Simplification des règles de la Figure 4.7.

Nous présentons aux médecins les règles de la Figure 4.8 pour prédire l'état de leurs patients. Nous avons obtenu uniquement six règles avec deux conjonctions, la prédiction des classes des patients est devenue très facile avec ces règles. L'exécution du même ensemble de données (WBC) en utilisant le nombre de partitions égale à 3 donne 27 règles avec deux conjonctions, le résultat reste toujours raisonnable. Si nous choisissons d'utiliser cinq règles avec trois partitions, nous aurons 243 règles. Cela explique notre choix du nombre de variables dans une combinaison.

4.5.3 Expérimentation

Afin d'expérimenter notre approche, nous allons calculer, comme critères de performances, l'exactitude, la précision, la sensitivité et la spécificité, en utilisant la matrice de confusion. Le Tableau 4.2 montre le pourcentage d'exactitude de notre approche en variant le nombre de partitions et le nombre de variables. Le seuil d'implifiance est égale à 0.8 pour les jeux de données WBC et WDBC. Il est égal à 0,5 pour le jeu de données WPBC dans le cas de deux partitions. Il est égal à 0,4 dans le cas de trois partitions. Nous n'avons pas utilisé la même valeur pour le seuil d'implifiance pour tous les jeux de données car, pour les jeux de données WBC et WDBC, nous avons plusieurs règles dont la valeur d'implifiance est très élevée, contrairement au jeu de données WPBC. Si nous exécutons ce dernier jeu de données avec le nombre de partitions égal à trois nous n'aurons que cinq variables dont la valeur d'implifiance est supérieur au seuil, comme on peut le voir au Tableau 4.2, avec le nombre de variables égal à 7, il n'y pas de résultat. C'est pour cela que nous avons mis NA. Donc nous devons diminuer la valeur d'implifiance afin d'avoir un nombre de règles suffisant.

L'exactitude obtenue avec les jeux de données WBC et WDBC est meilleur que celle obtenue avec WPBC. Cela parait très logique car la valeur d'implifiance utilisée avec les jeux de données

n-partition	n-variable	WBC	WDBC
2	3	86.61088	85.37736
3	3	94.97908	92.45283
3	5	96.65272	91.98113
3	7	96.23431	90.09434

Tableau 4.3 – Pourcentage de précision pour les ensembles de données WBC, WDBC.

n-partition	n-variable	WBC	WDBC
2	3	97.1831	96.79144
3	3	94.97908	92.45283
3	5	94.67213	95.12195
3	7	95.43568	95.5

Tableau 4.4 – Pourcentage de sensitivité pour les ensembles de données WBC, WDBC.

WBC et WDBC est très élevée en comparant à celle utilisée avec le jeu de données WPBC.

Nous pouvons avoir des résultats meilleurs si nous choisissons un nombre de variables plus élevé. Nous avons choisi d'utiliser uniquement trois variables car avec trois variables, le nombre de règles à présenter aux médecins pour prédire l'état de leurs patients est petit. En augmentant le nombre de variables, le nombre de règles devient considérablement plus grand.

Les Tableaux 4.3, 4.4 et 4.5 montrent respectivement le pourcentage de précision, le pourcentage de sensitivité et le pourcentage de spécificité de notre approche en variant le nombre de partitions et le nombre de variables pour les jeux de données WBC et WDBC. Le seuil d'implifiance utilisé est égal à 0,8 pour les deux ensembles de données.

4.5.4 Evaluation

Afin d'évaluer notre approche, nous la comparons aux différents classificateurs en Data Mining en utilisant les critères de performances précédemment calculés. Les Tableaux 4.6, 4.7 et 4.8 montrent les valeurs d'exactitude, de sensitivité, de spécificité de notre approche et d'autres outils en Data Mining. Les résultats de notre approche présentés dans ces tableaux sont ceux correspondant à trois partitions et à cinq règles.

Les résultats obtenus montrent que l'exactitude en utilisant notre approche est plus élevée que celle des autres approches pour les trois jeux de données (WBC, WDBC et WPBC). Le

n-partition	n-variable	WBC	WDBC
2	3	93.19149	91.88482
3	3	97.2973	95.51821
3	5	98.17768	95.32967
3	7	97.9638	94.30894

Tableau 4.5 – Pourcentage de spécificité pour les ensembles de données WBC, WDBC.

Algorithme	L'exactitude %		La sensitivité %	
	WBC	WDBC	WBC	WDBC
Le modèle Bayesien Naïf	96.50	92.61	95.7	89.6
RBF networks	96.66	93.67	95.9	90.0
Tree-J48	94.59	92.97	94.4	91.5
Trees-Cart	94.27	92.97	94.4	89.1
Notre approche	96.92	95.25	94.67	95.1

Tableau 4.6 – L'exactitude et la sensitivité des ensembles de données WBC et WDBC.

Algorithme	WPBC
Le modèle Bayesion Naïf	74.79
SMO	75.79
Tree-J48	74.74
Notre approche	77.27

Tableau 4.7 – L'exactitude de l'ensemble de données WPBC.

Algorithme	Spécificité %	
WBC	WDBC	
La méthode Bayésien Naïf	97.8	94.3
RBF networks	97.8	95.7
Tree-J48	92.6	93.8
Trees-Cart	93.9	95.2
Notre approche	98.17	95.32

Tableau 4.8 – Pourcentage de spécificité en utilisant les jeux de données WBC et WDBC.

pourcentage de sensitivité et de spécificité de notre approche est plus élevé ou très proche de celui des autres classificateurs. Les résultats obtenus par notre approche peuvent être plus intéressant en variant le nombre de partitions et aussi le nombre de règles choisies, comme le montrent les Tableaux 4.2, 4.3, 4.4 et 4.5. Par exemple l'exactitude de l'ensemble de données WBC obtenue peut atteindre 97.07 %. La sensitivité et la spécificité pour le même ensemble de données peut atteindre respectivement 97.18 % et 97.29 %. L'expert peut choisir une valeur faible ou élevée de ces paramètres. Nous proposons de choisir une valeur faible afin d'avoir peu de règles pour faire la prédiction. Finalement nous concluons que notre approche donne des résultats plus compréhensifs et elle est meilleur en termes de performance que les autres classificateurs pour les jeux de données WBC, WDBC et WPBC.

4.6 Approche 2 : Classification sans partitionnement de données

Dans cette approche nous proposons un nouveau classificateur basé sur les règles d'association, à la fois facile à interpréter, conduisant à une précision pertinente et permettant de mesurer la capacité de généralisation du modèle. A travers cette approche plusieurs améliorations ont été apportées à l'approche précédente.

Premièrement nous avons pensé à rendre l'approche de classification plus facile à utiliser, de telle façon qu'un simple utilisateur puisse : utiliser le code, trouver les mêmes résultats publiés, et plus encore, changer les paramètres librement et trouver d'autres solutions. Par la suite nous avons fait en sorte qu'il soit possible d'exécuter notre code en dehors de RCHIC. Pour cela nous avons reproduit les mesures d'intérêt qui nous intéresse, telles que l'implifiance, car dans l'ancienne approche ces valeurs sont automatiquement calculées par le logiciel RCHIC dans le fichier transaction.out.

Dans le but d'avoir des résultats plus précis nous avons raisonné d'une manière différente pour faire les prédictions. Dans cette nouvelle approche chaque individu (ligne) est classé en fonction des valeurs de chacune des variables significatives sélectionnées (colonne) contrairement à la première approche où les prédictions se font selon l'échantillon où appartient la variable de l'individu dont on veut prédire. Dans ce cas nous avons éliminé le partitionnement des données.

Dans cette approche nous avons également reproduit les résultats des algorithmes de DATA Mining, et cela à travers le package RWeka, ce qui nous a permis d'exécuter tous les algorithmes y compris notre approche dans les mêmes conditions. Cela rend la comparaison et l'évaluation équitable.

Pour tester l'efficacité de notre approche, nous avons utilisé la validation croisée "cross-validation" sous sa forme la plus populaire qu'est le k-fold cross-validation. Ce qui nous a permis de comparer notre méthode aux différents modèles de ML.

Nous avons également testé notre approche par de nouveaux jeux de données (Haberman, SOBAR et IDs-Mapping), en plus des jeux de données sur le cancer de sein (WBC, WDBC et WPBC).

4.6.1 Description de l'approche

Cette nouvelle méthode de classification, utilise les règles d'implication de l'ASI pour faire les prédictions mais avec un raisonnement un peu différent de l'approche précédente et avec beaucoup d'améliorations.

Avant toutes étapes de traitement des données, notre approche effectue la cross-validation, où l'ensemble d'apprentissage et l'ensemble de test seront générés à partir de l'ensemble de données initial et cela en divisant aléatoirement l'ensemble de données en k parties (5). Nous considérons

une partie comme l'ensemble de test avec lequel nous allons évaluer notre approche et l'union des k-1 (4) autres parties comme ensemble d'apprentissage selon lequel nous avons construit notre classificateur.

Le classificateur commentce d'abord à générer toutes les règles d'implications et cela en faisant appel à l'algorithme Apriori. Par la suite, toutes les règles possibles pour lesquelles la conclusion représente un nom d'une classe (toutes les règles avec lesquelles nous pouvons faire la prédiction) seront sélectionnées. La variable de classification 'classe' est dupliquée en variables binaires classe0 et classe1. Nous avons exploité toutes les règles de classification de longueur 2 (dont le côté droit est soit classe0, soit classe1). Parmi ces règles, nous sélectionnons uniquement les règles les plus importantes. Pour cela un seuil d'implifiance (le premier paramètre) a été défini, puis les règles qui dépassent cette implifiance seront filtrées. La valeur d'implifiance a été calculée en utilisant l'équation (1.2) du Chapitre 1. Les sous-ensembles de variables significatives seront déduit a partir des règles sélectionnées. Les variables ayant un faible effet sur la variable de classification classe sont rejetées.

Le choix des variables significatives se fait selon le seuil d'implifiance choisi. Par exemple, si une variable V possède 3 valeurs a, b et c on considère alors les règles où chaque valeur implique la classe positive ($a \Rightarrow 1, b \Rightarrow 1 \ et \ c \Rightarrow 1$), et les règles où chaque valeur implique la classe négative ($a \Rightarrow 0, b \Rightarrow 0 \ et \ c \Rightarrow 0$). Les règles avec des valeurs d'implifiance supérieur au seuil seront retenues. Si pour la valeur a, les deux règles qui la composent ne sont pas significatives, il ne sera pas possible d'utiliser cette variable V pour classer un individu pour lequel V = a. Alors V n'est pas une variable significative.

Le nombre de variables à utiliser pour faire les prédictions est sélectionné par l'utilisateur, il s'agit d'un nombre impair faible « nvote » de "prémisses" (un autre paramètre). Nous choisissons un nombre faible car plus le nombre de variables utilisées pour définir le processus de classification est élevé, plus l'erreur de prédiction est faible. A l'inverse, la méthode de classification résultante sera moins compréhensible pour l'utilisateur.

Dans cette approche, toutes les combinaisons de taille 1 à nvote variables seront formées (plusieurs niveaux de prédictions). Contrairement a la première approche qui prend en considération que les combinaisons composées de nvote variables (un seul niveau de prédiction). Donc l'étape suivante consiste à établir les tableaux des classifications de 1 à nvote prédicteurs (prédiction avec une seule règle, jusqu'à une combinaison contenant nvote règles). les combinaisons sont formées en utilisant les règles significatives.

Chaque individu est classé en fonction des valeurs de chacune des variables significatives sélectionnées (voie l'exemple d'application), contrairement à la première approche où la classification d'un individu dépend de l'appartenance de ces variables aux échantillons formés.

Durant le processus de classification, la prédiction de chaque individu avec chaque combinaison de variable est réaliser par un vote majoritaire.

Le groupe final de variables significatives (la combinaison de variable utilisée pour la prédiction) est celui dont l'erreur de prédiction est la plus faible (parmi toutes les combinaisons dans tous les niveaux de prédiction de 1 à nvote). Ce dernier sera conforme à la méthode de classification finale.

La répartition des données entre l'ensemble d'apprentissage et l'ensemble de test est aléatoire. Lors de la tache de prédiction (dans l'ensemble de test) certaines valeurs peuvent ne pas être classifié avec les anciennes règles générées, car ces valeurs peuvent ne pas exister dans l'ensemble d'apprentissage avec lequel nous avons conçu notre classificateur (Ces valeurs de NA apparaissent dans la section d'évaluation pour les autres algorithmes de classification). Dans notre approche nous avons traité ces cas. Nous avons attribué à la valeur qui donne le NA, la prédiction de la valeur la plus proche à cette dernière. Par exemple, si la valeur qui donne le NA est inférieure ou supérieure à toutes les valeurs dans l'ensemble d'apprentissage, on leur attribue respectivement la prédiction de la plus grande et la plus petite valeur dans l'ensemble d'apprentissage.

4.6.2 Implémentation de l'approche

Nous avons implémenté l'algorithme dans le logiciel R sous Linux, disponible gratuitement dans [Gha22]. Il contient la fonction siacrossvalid(), dans laquelle nous avons appliqué la cross validation à notre approche. Il contient également la fonction SIAclassif(), qui développe l'algorithme et renvoie le classificateur, ainsi que la fonction predict.SIA(), qui prend l'objet donné par la fonction précédente et l'applique à de nouvelles données.

La fonction siacrossvalid() consiste à appliquer la cross validation à notre approche et aux autres algorithmes de machine learning avec lesquels nous l'avons comparé. Pour ce faire nous avons généré l'ensemble d'apprentissage et l'ensemble de test à partir de l'ensemble de données initial. L'ensemble d'apprentissage est utilisé pour réaliser le classificateur et l'ensemble de test pour l'évaluer.

La fonction SIAclassif() est celle qui réalise toutes les étapes de notre algorithme, en fonction de l'implifiance et du nombre maximum de nvotes. Elle calcule l'implifiance de toutes les règles, elle sélectionne les variables significatives, et elle calcule toutes les combinaisons de groupes de variables significatives et la classification résultante pour chacune d'elles (1 prédicteur jusqu'à nvote prédicteurs), en prenant celle qui maximise la précision (par défaut, ou tout autre critère spécifié par l'utilisateur). Notre classificateur renvoi un test expliquant les règles de classification. À titre d'exemple, le résultat pour le jeu de données WBC est présenté dans la section suivante.

La fonction predict.SIA() requiert, comme premier argument, l'objet retourné par la fonction SIAclassif(), et comme second argument, un ensemble de données avec les nouvelles instances(l'ensemble de test), afin d'appliquer la classification et de produire la sortie prédite. Elle traite également les valeurs qui donnent les NA lors de la tache de prédiction des instances figurant dans l'ensemble de test est pas d'apprentissage.

Dans la section suivante, on trouve un exemple de résultat donné par notre algorithme, qui

explique à l'utilisateur comment les instances sont classées en fonction des valeurs des variables significatives.

4.6.3 Exemple d'application

A titre d'exemple pour les utilisateurs, nous montrons dans cette section la sortie de notre classificateur. Le résultat de la fonction SIAclassif() est stockée dans un objet appelé « rules ». Nous prenons comme exemple l'ensemble de données WBC, en utilisant jusqu'à 7 variables significatives. Les classes de sortie sont "malignes" et "bénignes", tandis que les caractéristiques d'entrée sont $V_1, V_2, ..., V_8$ *et* V_9. Notre classificateur offre à l'utilisateur la possibilité de choisir le fichier de données et le paramètre *nvote* à travers une interface conviviale simple.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	class
1	5	1	1	1	2	1	3	1	1	benign
2	5	4	4	5	7	10	3	2	1	benign
3	3	1	1	1	2	2	3	1	1	benign
4	6	8	8	1	3	4	3	7	1	benign
5	4	1	1	3	2	1	3	1	1	benign
6	8	10	10	8	7	10	9	7	1	malignant
7	1	1	1	1	2	10	3	1	1	benign
8	2	1	2	1	2	1	3	1	1	benign
9	2	1	1	1	2	1	1	1	5	benign
10	4	2	1	1	2	1	2	1	1	benign
11	1	1	1	1	1	1	3	1	1	benign

FIGURE 4.9 – Extrait du jeu de donnée WBC.

1. Si l'utilisateur souhaite utiliser une seule variable, la meilleure classification est réalisée en utilisant la variable V_2, qui prédit la classe "maligne" pour un individu chaque fois qu'il prend l'une des valeurs 10, 3, 4, 5, 6, 7, 8 ou 9. La précision de ce classificateur est de 0.926793557833089.
2. Si l'utilisateur accepte d'utiliser jusqu'à trois variables ($nvote = 3$), la meilleure classification atteint une précision de 0.961932650073206 en vérifiant les variables V_8, V_3 *et* V_6 ou bien la combinaison V_7, V_3 *et* V_6. La règle prévoit la classe "maligne" pour un enregistrement, si au moins deux des variables V_8, V_3 *et* V_6 prennent une de leurs valeurs dans les ensembles suivant : $V_8 = 10, 3, 4, 5, 6, 7, 8, 9$; $V_3 = 10, 3, 4, 5, 6, 7, 8, 9$; $V_6 = 10, 3, 4, 5, 6, 7, 8, 9$ ou bien si au moins deux des variables V_7, V_3 *et* V_6 prennent une de leurs valeurs dans les ensembles : $V_7 = 10, 4, 5, 6, 7, 8, 9$; $V_3 = 10, 3, 4, 5, 6, 7, 8, 9$; $V_6 = 10, 3, 4, 5, 6, 7, 8, 9$.

3. Si l'utilisateur choisi d'utiliser 5 variables significatives ($nvote = 5$), la meilleure classification atteint une précision de 0.972181551976574 et s'il choisi 7 variables significatif, la meilleur classification atteint une précision de 0.97510980966325. A chaque fois que le nombre de variable significatives augmente la précision du modèle devient meilleurs. (Pour plus de détail voire la sortie d'exécution)
4. En général, l'utilisateur choisit un classificateur particulier en fonction de l'importance accordée à la précision, ainsi qu'à la simplicité de la règle. Dans tous les cas, l'interprétation de la classification de notre approche est facile, contrairement à d'autres classificateurs, car le praticien n'a besoin de vérifier que les valeurs de quelques variables originales pour décider de la classe de sortie. L'utilisateur peut choisir de maximiser d'autres critères tels que la précision, la sensitivité, la spécificité.

Le (Tableau 4.9) représente la sortie de notre classificateur quant l'utilisateur tape «rules» dans la console R.

4.6.4 Résultats et discussions

Afin d'évaluer notre approche, nous l'avons comparé à d'autres algorithmes largement utilisées dans la littérature, avec des ensembles de données en accès libre provenant du dépôt Machine Learning de l'UCI. Nous allons calculer, comme critères de performances, l'exactitude, la précision, la sensitivité et la spécificité, en utilisant la matrice de confusion.

Nous avons implémenté les différents algorithmes de DM utilisés pour l'évaluation en utilisant le package Rweka dans le logiciel R. Comme indiqué dans la section précédente, nous avons utilisé la cross validation pour un nombre de partie k=5. Les jeux de données ont été divisés aléatoirement en cinq groupes. En fixant l'un de ces groupes, toutes les méthodes seront exécutées par tous les autres groupes d'instances. Ensuite, à partir des méthodes exécutées, les instances du groupe qu'on a fixé seront classifiées, et la matrice de confusion ainsi que les quatre critères de performance seront calculés.

En répétant le processus pour chacun des cinq groupes et en faisant la moyenne des valeurs obtenues, on obtient les résultats présentés dans le Tableau 4.10 et représentés dans la Figure 4.10 pour faciliter l'interprétation.

Le Tableau 4.10 montre les valeurs des quatre mesures de performances utilisés pour évaluer notre approche et les autres algorithmes et cela pour chaque jeu de données. la Figure 4.10 montre une représentation graphique pour faciliter la lecture et la comparaison des résultats.

Lorsque les classes ne sont pas équilibrées, certaines classes peuvent ne pas apparaître dans certains ensembles de formation et de test, ce qui entraîne des valeurs indéfinies des critères de performance (à cause de la division par 0). Ce cas ne peut pas figurer dans notre approche car nous l'avons traité (dans la fonction predict.SIA()).

```
[1] "rules"
[[1]]
[1] "Classify as 'malignant' iff at least 1 of the following variables hold:
V2={10,3,4,5,6,7,8,9}; . The 'Exactitude' is 0.926793557833089."

[[2]]
[1] "Classify as 'malignant' iff at least 2 of the following variables hold:
V8={10,3,4,5,6,7,8,9}; V3={10,3,4,5,6,7,8,9}; V6={10,3,4,5,6,7,8,9}; .
The 'accuracy' is 0.961932650073206."

[[3]]
[1] "Classify as 'malignant' iff at least 2 of the following variables hold:
V7={10,4,5,6,7,8,9}; V3={10,3,4,5,6,7,8,9}; V6={10,3,4,5,6,7,8,9}; .
The 'accuracy' is 0.961932650073206."

[[4]]
[1] "Classify as 'malignant' iff at least 3 of the following variables hold:
V8={10,3,4,5,6,7,8,9}; V7={10,4,5,6,7,8,9}; V3={10,3,4,5,6,7,8,9};
V6={10,3,4,5,6,7,8,9}; V1={10,6,7,8,9}; . The 'accuracy' is 0.972181551976574."

[[5]]
[1] "Classify as 'malignant' iff at least 3 of the following variables hold:
V7={10,4,5,6,7,8,9}; V5={10,3,4,5,6,7,8,9}; V2={10,3,4,5,6,7,8,9};
V6={10,3,4,5,6,7,8,9}; V1={10,6,7,8,9}; . The 'accuracy' is 0.972181551976574."

[[6]]
[1] "Classify as 'malignant' iff at least 3 of the following variables hold:
V7={10,4,5,6,7,8,9}; V5={10,3,4,5,6,7,8,9}; V3={10,3,4,5,6,7,8,9};
V6={10,3,4,5,6,7,8,9}; V1={10,6,7,8,9}; . The 'accuracy' is 0.972181551976574."

[[7]]
[1] "Classify as 'malignant' iff at least 3 of the following variables hold:
 V7={10,4,5,6,7,8,9}; V2={10,3,4,5,6,7,8,9}; V3={10,3,4,5,6,7,8,9};
 V6={10,3,4,5,6,7,8,9}; V1={10,6,7,8,9}; . The 'accuracy' is 0.972181551976574."

[[8]]
[1] "Classify as 'malignant' iff at least 4 of the following variables hold:
V8={10,3,4,5,6,7,8,9}; V7={10,4,5,6,7,8,9}; V5={10,3,4,5,6,7,8,9};
V2={10,3,4,5,6,7,8,9}; V3={10,3,4,5,6,7,8,9}; V6={10,3,4,5,6,7,8,9};
V1={10,6,7,8,9}; . The 'accuracy' is 0.97510980966325."
[[9]]
[1] "Classify as 'malignant' iff at least 4 of the following variables hold:
V8={10,3,4,5,6,7,8,9}; V7={10,4,5,6,7,8,9}; V5={10,3,4,5,6,7,8,9};
V3={10,3,4,5,6,7,8,9}; V6={10,3,4,5,6,7,8,9}; V1={10,6,7,8,9};
V4={10,2,3,4,5,6,7,8,9}; . The 'accuracy' is 0.97510980966325."
```

Tableau 4.9 – Exemple d'un résultat du classificateur.

Comme le montre la Figure 4.10, notre proposition obtient une première ou une deuxième position parmi les cinq classificateurs dont quatre des six ensembles de données, en ce qui concerne l'exactitude, la précision et la sensitivité.

Le principal avantage de notre approche est la simplicité d'interprétation de la règle de classification (voir l'exemple d'application), puisque la classification est directement liée aux variables observées.

On peut considérer que, dans le compromis entre précision et interprétabilité du modèle, notre proposition dépasse toutes les autres méthodes testées. Même si J48 et CART sont également interprétables, leurs arbres sont généralement plus complexes que les règles que nous obtenons avec notre méthode.

Dataset	Method	Exactitude	Precision	Sensitivity	Specificity
WDBC	NB	0.7240773	0.6629947	0.5329353	0.8361464
	CART	0.6432337	0.5986015	0.1629407	0.9303413
	J48	0.6274165	NA	0.0000000	1.0000000
	RadSVM	0.6274165	NA	0.0000000	1.0000000
	SIA	0.8347979	0.7786700	0.7688728	0.8728805
IDs_mapping	NB	0.6417223	0.6092675	0.5617475	0.7067785
	CART	0.6578772	0.6288379	0.5829920	0.7187949
	J48	0.6104806	0.6032557	0.3858690	0.7933888
	RadSVM	0.5512016	NA	0.0000000	1.0000000
	SIA	0.6550734	0.6325261	0.5529404	0.7383278
SOBAR	NB	0.9027778	0.9652778	0.7460317	0.9739583
	CART	0.6527778	0.5027778	0.4365079	0.7619464
	J48	0.7083333	NA	0.0000000	1.0000000
	RadSVM	0.7083333	NA	0.0000000	1.0000000
	SIA	0.8055556	0.7833333	0.7301587	0.8952020
WBC	NB	0.9765739	0.9475863	0.9870416	0.9707784
	CART	0.9458272	0.9225932	0.9247400	0.9574295
	J48	0.9311859	0.9130195	0.8921389	0.9536268
	RadSVM	0.9663250	0.9570372	0.9447121	0.9775771
	SIA	0.9428990	0.9107039	0.9285842	0.9500720
WPBC	NB	0.6752577	0.3009450	0.2740059	0.7963057
	CART	0.7577320	0.5552283	0.1666237	0.9403461
	J48	0.7628866	NA	0.0000000	1.0000000
	RadSVM	0.7628866	NA	0.0000000	1.0000000
	SIA	0.6752577	0.2424399	0.1640648	0.8377091
haberman	NB	0.7156863	0.4185765	0.2281000	0.8850047
	CART	0.6960784	0.2857298	0.2097401	0.8622492
	J48	0.7352941	NA	0.0000000	1.0000000
	RadSVM	0.7352941	NA	0.0000000	1.0000000
	SIA	0.6143791	0.2898471	0.2910234	0.7110647

Tableau 4.10 – Résultats de classification pour notre classificateur et les méthodes ML appliquées à chaque ensemble de données.

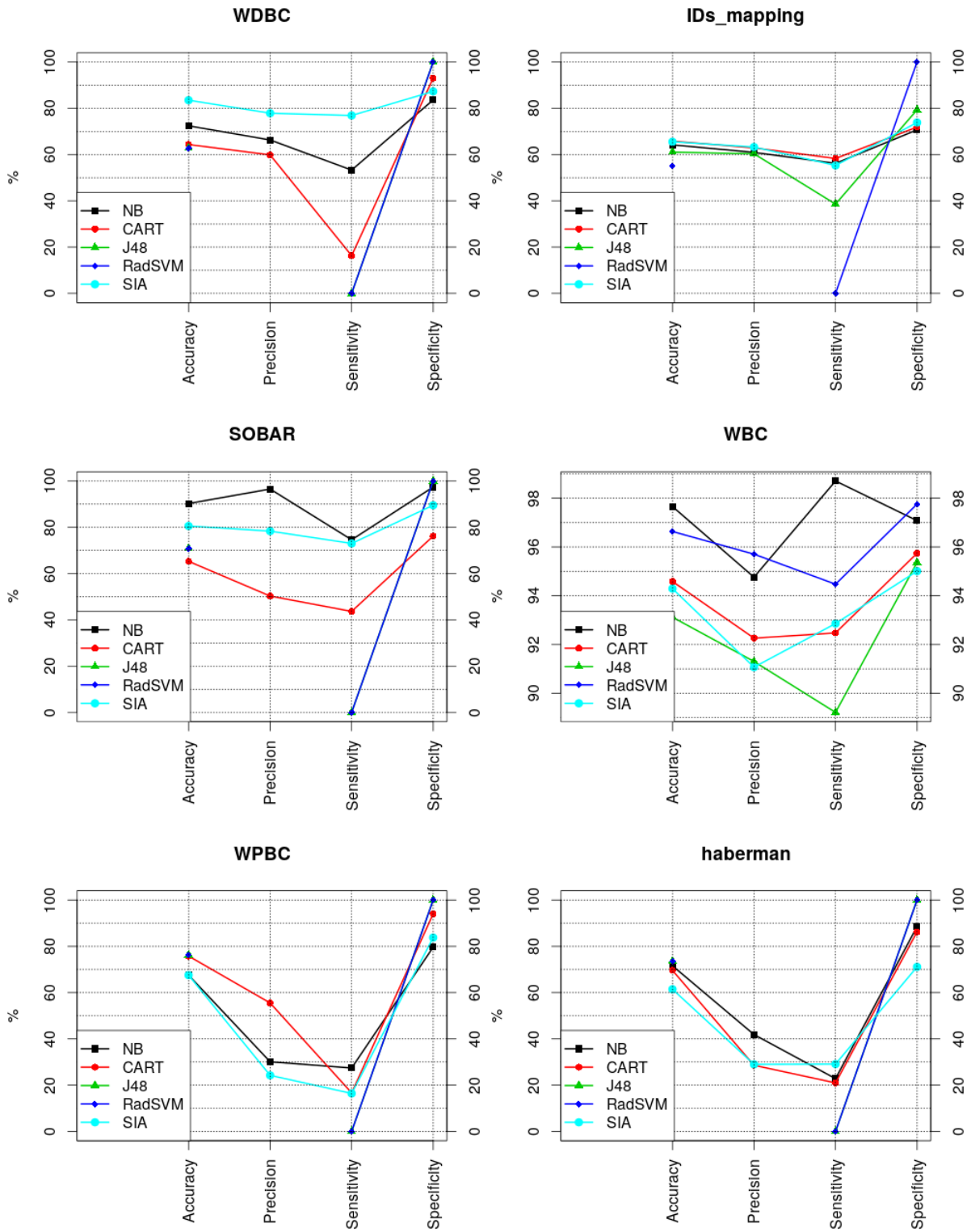

FIGURE 4.10 – Représentation graphique du Tableau 4.10.

4.7 Conclusion

Dans ce chapitre nous avons proposé deux algorithmes d'apprentissage automatique pour la classification basée sur les règles d'association (les règles du graphe d'implication de l'ASI) et un vote majoritaire parmi un ensemble de règles significatives.

Dans la première approche les variables sont partitionnées. Chaque partition désigne une classe donnée. La classification d'un individu dépend de l'appartenance de ces variables aux échantillons formés. La variable classe pour cet individu correspond a la classe majoritaire de l'ensemble des classes données par chaque échantillon où appartient chaque variable de ce dernier. Contrairement à la deuxième approche où chaque individu est classé en fonction des valeurs de chacune des variables significatives sélectionnée (les variables ne sont pas partitionnées)

Pour évaluer la première approche, nous avons utilisé les critères de performances suivant : l'exactitude, la précision, la sensitivité et la spécificité, avec des ensembles de données sur le cancer de sein. Nous avons comparé notre approche à d'autres approches du Data Mining, telles que la méthode naïve bayésienne, les réseaux de neurones à base radiale, les arbres de décisions J48 et simple CART. Notre approche peut souvent être meilleure que les autres classificateurs en Data Mining en termes de performances pour les jeux de données utilisés. Dans le domaine de la médecine, notre approche aide les médecins dans leur prise de décision. Elle permet d'améliorer les résultats des patients, réduire le coût de la médecine et aider à améliorer les études cliniques.

La deuxième approche a été testée à l'aide de six ensembles de données en libre accès (en plus des jeux de données sur le cancer de sein, deux autres on été utilisés : Haberman et SOBAR). Elle a été comparée aux quatre algorithmes de classification utilisés dans l'approche 1, et a été évalué avec les mêmes critères de performance utilisé dans l'approche 1, plus la validation croisée à cinq reprises. D'après les résultats obtenus notre proposition obtient une première ou une deuxième position parmi les cinq classificateurs dont quatre des six ensembles de données, en ce qui concerne l'exactitude, la précision et la sensitivité.

Le principal avantage des deux approches proposées est la simplicité d'interprétation de la règle de classification (voir les exemples d'application Tableau 4.9 et Figure 4.8). Nos approches donnent des résultats qui sont compréhensibles pour le simple utilisateur, car la démarche suivie est très logique, du fait que la classification est directement liée aux variables observées.

Conclusion générale

La fouille de données est réalisée automatiquement par des algorithmes combinatoires, la fouille de règles est généralement laissée à la charge de l'utilisateur. En pratique, il est très laborieux pour ce dernier de rechercher des connaissances intéressantes dans les listes de règles obtenues à la sortie des algorithmes. Différentes solutions ont été proposées pour aider l'utilisateur à réaliser cette tâche. Dans cette thèse nous nous sommes intéressés à une approche statistique originale dédiée à l'extraction et à l'analyse de règles. Il s'agit de l'Analyse Statistique Implicative (ASI). Elle est basée sur la mesure de l'intensité d'implication qui mesure l'étonnement d'observer un petit nombre de contre-exemples à une règle. Ce raisonnement a été affirmé par René Thom dans « Paraboles et catastrophes », 1980, p.130 : «...le problème n'est pas de décrire la réalité, le problème consiste bien plus à repérer en elle ce qui a de sens pour nous, ce qui est surprenant dans l'ensemble des faits. Si les faits ne nous surprennent pas, ils n'apportent aucun élément nouveau pour la compréhension de l'univers : autant donc les ignorer » [Ren80].

Notre premier objectif de la thèse consiste à améliorer et à créer certaines fonctions permettant de calculer les méthodes de l'ASI dans R. Il s'agit de poursuivre des travaux entrepris à ce sujet. De nombreux travaux existent sur l'ASI mais en dehors de R. Le fait d'intégrer les travaux majeurs permettra à la communauté de bénéficier des avantages de R (simplicité, portabilité, reconnaissance) et permettra également d'améliorer la diffusion de l'ASI. Par la suite nous avons introduit d'autres fonctionnalités à l'ASI. Nous avons rajouté un nouveau mode pour la sélection des règles dans le graphe d'implication qui offre à l'utilisateur la possibilité de sélectionner non seulement les règles les plus étonnantes mais également celles dont le pourcentage de participation est très élevé. Nous avons offert la possibilité à l'utilisateur de faire de la classification avec l'ASI et cela à travers l'introduction et le développement de deux classificateurs basés sur l'ASI.

L'ASI est une branche de recherche peu connu en Algérie, cela nous a motivé a fixé un but initial qui consiste à montrer l'intérêt de l'ASI et susciter son intérêt pour les administrations, les laboratoires de recherches et les différentes structures pour avoir des perspectives et des prévisions afin d'optimiser leurs résultats. La théorie ASI pourrait paraître compliqué aux utilisateurs non statisticien et non informaticien. Pour palier ces problèmes, notre première contribution est un pré-requis sur les concepts de base pour qu'un utilisateur en ASI puisse assimiler l'apport de l'ASI, avec un état de l'art sur la théorie ASI, (R)CHIC et sur quelques méthodes de classification en

Data Mining.

Dans la deuxième contribution nous avons étudié les notes des étudiants de l'Université de Béjaia, cela nous a permis de nous familiariser avec le logiciel conçu pour l'ASI, de fournir des informations scientifiques utiles pour les futurs étudiants, telles que l'orientation des étudiants vers une spécialité donnée, avoir une vision sur les causes d'échecs dans certains modules. Cela à travers l'organisation, la structuration et l'extraction des différentes liaisons entre les modules étudiés, en utilisant le graphe implicatif du logiciel CHIC et l'analyse des résultats obtenus. Cette contribution a fait l'objet de la conférence internationale ISKO-Maghreb 2014 .

Dans la troisième contribution nous avons étudié le couplage de l'indice d'implication avec la confiance, afin d'améliorer et de faciliter la recherche de règle avec la méthode d'analyse statistique implicative. Cela a été réalisé à travers l'ajout d'un nouveau mode dans le menu RCHIC (intensité d'implication+confiance). Avec ce nouveau mode la valeur de la confiance est affichée pour chaque règle dans le graphe, et un seuil de confiance a été ajouté afin de permettre l'affichage des règles qui sont supérieur à ce seuil. Cela permet à l'utilisateur d'afficher que les règles étonnantes auxquelles il peut faire confiance et le graphe devient plus lisible. Ce travail a été présenté dans la conférence ASI8 en 2015.

Dans la quatrième contribution nous avons proposé un nouveau classificateur basé sur les règles d'implication de l'ASI, qui utilisent comme mesure de qualité l'implifiance. Dans cette méthode, les données sont partitionnées en utilisant l'algorithme des nuées dynamiques. Par la suite l'algorithme apriori a été utilisé pour générer les règles d'implication à partir des classes existantes et des échantillons formés. Notre classificateur ne garde que les règles importantes. A partir de ces règles les combinaisons de variables seront formées. Le nombre de variables dans la combinaison est choisi par l'utilisateur. L'erreur est calculée pour chaque combinaison, et celle ayant la plus petite erreur de prédiction est utilisée pour la classification. L'approche a été expérimentée sur des ensembles de données sur le cancer de sein en libre accès, et ses résultats ont été comparés à quatre autres algorithmes de classification bien établis (Naive Bayes, Réseaux de neurones à base radiale, Arbres de décision J48 et CART simple). Ce travail a été publié à la conférence international ASI9 en 2017.

Dans la dernière contribution nous avons réalisé un autre classificateur à base de règles d'implication de l'ASI. En plus de sa facilité d'interprétation et de sa précision nous avons exécuté l'approche en dehors de Rchic, pour qu'un simple utilisateur puisse : installer le package R, utiliser le code, trouver les mêmes résultats publiés, et plus encore, changer les paramètres librement et trouver d'autres solutions. Dans cette méthode, nous avons éliminé le partitionnement des données et l'algorithme apriori génère les règles d'implication à partir des classes et variables existantes. Notre classificateur garde que les règles importantes. A partir de ces règles les combinaisons de variables seront formées. Le nombre de variables dans la combinaison est choisi par l'utilisateur. L'erreur est calculée pour chaque combinaison, et celle ayant la plus petite erreur de prédiction

est utilisée pour la classification. Chaque individu est classé en fonction des valeurs de chacune des variables significatives dans la combinaison ayant l'erreur de prédiction la plus faible. Dans cette approche nous avons reproduit les résultats des algorithmes de DM, et cela à travers le package RWeka, et nous avons appliqué la cross validation à notre approche et à toutes les autres approches, ce qui nous a permis d'exécuter tous les algorithmes et notre approche dans les mêmes conditions. Cela rend la comparaison et l'évaluation équitable. En plus des jeux de données sur le cancer de sein (WBC, WDBC et WPBC), le classificateur a été également testé par de nouveaux jeux de données (Haberman, SOBAR et IDs Mapping). Les résultats obtenus ont été comparés aux résultats obtenus des autres algorithmes de classification (Naive Bayes, Réseaux de neurones à base radiale, Arbres de décision J48 et CART simple). Dans le compromis entre la précision et la facilité d'interprétation du modèle, notre classificateur surpasse toutes les autres méthodes testées. Même si J48 et CART sont également interprétables, leurs arbres sont généralement plus complexes que les règles que nous avons obtenu avec notre classificateur. Ce travail a été publié dans le journal Mathematics en 2021.

A l'issu de ce travail, de nombreuses perspectives s'ouvrent sur les différents points traités. Les principales sont :

En plus de l'analyse implicative effectuée sur les notes des étudiants, nous proposons d'essayer d'autres possibilités qu'offre l'ASI telles que l'utilisation du graphe cohésif qui permet de voir les niveaux des différentes implications.

Nous envisageons d'utiliser d'autres mesures que la confiance parmi celles disponibles dans [GH07]. Cela va permettre aux utilisateurs de bénéficier des avantages de l'ASI et des différentes mesures déjà existantes. Les combinaisons formées vont permettre d'enrichir l'ASI.

Les deux approches de classification proposées sont faciles à interpréter, c'était l'objectif visé dès le départ. Les deux classificateurs proposent des résultats meilleurs ou proches des autres classificateurs en Data Mining. Les résultats obtenus par le premier classificateur sont meilleurs par rapport au deuxième. Pour cela nous proposons d'effectuer des améliorations à la première approche de classification. Nous gardons le partitionnement des données où chaque variable est partitionnée en un nombre fixe d'intervalles. Les prédictions se font selon l'échantillon qui contient la variable de l'individu dont on veut prédire. Nous proposons de reproduire les résultats des algorithmes de DATA Mining, avec le package RWeka, et d'appliquer la cross validation à ces algorithmes ainsi qu'à notre classificateur, comme on l'a fait avec le deuxième classificateur. Cela va nous permettre d'exécuter tous les algorithmes dans les mêmes conditions et d'avoir une comparaison équitable. Nous proposons aussi de tester cette approche avec de nouveaux jeux de données en plus des jeux de données sur le cancer de sein.

Pour le deuxième classificateur proposé, les paramètres par défaut ont été conservés. L'analyse de l'effet du choix de valeurs différentes aux paramètres dans chaque modèle va porter des améliorations a cette approche.

Bibliographie

[AAH14] N. Abdelhamid, A. Ayesh, and W. Hadi. Multi-label rules algorithm based associative classification. *Parallel Processing Letters*, 24(01) :1450001, 2014.

[AAT15] N. Abdelhamid, A. Ayesh, and F. Thabtah. Emerging trends in associative classification data mining. *International Journal of Electronics and Electrical Engineering*, 3(1) :50–53, 2015.

[AB94] DG. Altman and DJ. Bland. Diagnostic tests. 1 : Sensitivity and specificity. *BMJ : British Medical Journal*, 308(6943) :1552, 1994.

[AIS93] R. Agrawal, T. Imielienski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, 1993.

[Ale71] V. Alexandre. *Les echelles d'attitude : pref. d'abraham A. Moles*. Editions universitaires, 1971.

[AN07] A. Asuncion and D. Newman. Uci machine learning repository, 2007.

[AR+11] S. Aruna, SP. Rajagopalan, et al. Knowledge based analysis of various statistical tools in detecting breast cancer. *Computer Science & Information Technology*, 2(2011) :37–45, 2011.

[AS+94] R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499. Citeseer, 1994.

[Bay00] JLL. Del Bayle. *Initiation aux méthodes des sciences sociales*. L'harmattan, 2000.

[BBT10] M. Bailleul, JY. Bodergat, and JF. Themines. Logiques et significations des collectifs pour les enseignants débutants. *Education & Formation*, 2010.

[Bel08a] S. Belciug. Bayesian classification vs. k-nearest neighbour classification for the non-invasive hepatic cancer detection. In *Proc. 8th International conference on Artificial Intelligence and Digital Communications*, 2008.

[Bel08b] S. Belciug. Bayesian classification vs. k-nearest neighbour classification for the non-invasive hepatic cancer detection. In *Proc. 8th International conference on Artificial Intelligence and Digital Communications*, 2008.

[BFOS84] L. Breiman, JH. Friedman, RA. Olshen, and CJ. Stone. Classification and regression trees. wadsworth. *Inc. Monterey, California, USA*, 1984.

[Bla05] J. Blanchard. *Un système de visualisation pour l'extraction, l'évaluation, et l'exploration interactives des règles d'association.* PhD thesis, Université de Nantes, 2005.

[BMS97a] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets : Generalizing association rules to correlations. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, pages 265–276, 1997.

[BMS97b] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets : Generalizing association rules to correlations. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, pages 265–276, 1997.

[BSGG04] H. Briand, M. Sebag, R. Gras, and F. Guillet. Mesures de qualité pour la fouille de données, revue des nouvelles technologies de l'information (rnti) numéro spécial, 2004.

[CA09] R. Couturier and SA. Almouloud. Historique et fonctionnalités de chic., 2009.

[CCA96] S. Chen, ES. Chng, and K. Alkadhimi. Regularized orthogonal least squares algorithm for constructing radial basis function networks. *International journal of control*, 64(5) :829–837, 1996.

[CCST98] N. Cristianini, C. Campbell, and J. Shawe-Taylor. Dynamically adapting kernels in support vector machines. *Advances in neural information processing systems*, 11, 1998.

[CD89] M. Crahay and A. Delhaxhe. La compréhension du fonctionnement de la balance : Une analyse hiérarchique. *European Journal of Psychology of Education*, 4 :349–375, 1989.

[CG05] R. Couturier and R. Gras. Chic : traitement de données avec l'analyse implicative. In *EGC*, pages 679–684, 2005.

[Cou00] R. Couturier. Traitements de l'analyse implicative avec chic. *Journées sur l'implication statistique*, pages 33–50, 2000.

[Cou01] R. Couturier. Traitement de l'analyse statistique implicative dans chic. *Actes des Journées sur la Fouille dans les données par la méthode d'analyse implicative, IUFM Caen*, pages 33–50, 2001.

[Cou07] R. Couturier. Nouveaux apports théoriques à l'analyse statistique implicative et applications. chic : Utilisation et fonctionnalités. p. 41-49. 2007.

[Cou08] R. Couturier. Chic : Cohesive hierarchical implicative classification. In *Statistical implicative analysis*, pages 41–53. Springer, 2008.

[Cou15] R. Couturier. Un système de recommandation basé sur l'analyse statistique implicative. In *Actes du 8éme Colloque International sur Analyse statistique Implicative, Radés Tunisie*, pages 447–452, 2015.

[Cou22] R. Couturier. https ://members.femto-st.fr/raphael-couturier/en/rchic, (last update 01/03/2022).

[CV95] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3) :273–297, 1995.

[Dem04] M. Demeuse. introduction aux théories et aux méthodes de la mesure en sciences psychologiques et en sciences de l'éducation, 2004.

[DGB+08] St. Daviet, F. Guillet, H. Briand, S. Baquedano, V. Philippé, and R. Gras. Using the statistical implicative analysis for elaborating behavioral referentials. In *Statistical Implicative Analysis*, pages 299–319. Springer, 2008.

[Did71] E. Diday. Une nouvelle méthode en classification automatique et reconnaissance des formes la méthode des nuées dynamiques. *Revue de statistique appliquée*, 19(2) :19–33, 1971.

[DR17] D. Diaz and JC. Regnier. Étude des difficultés et facilités d'apprentissage de la statistique à la lumière de l'asi, 2017.

[EP96] J. Elder and D. Pregibon. A statistical perspective on kdd. *Advances in knowledge discovery and data mining*, pages 83–116, 1996.

[FGH+12] A. Fink, R. German, M. Heron, S. Stewart, C. Johnson, J. Finch, D. Yin, P. Schaeffer, Accuracy of Cancer Mortality Working Group, et al. Impact of using multiple causes of death codes to compute site-specific, death certificate-based cancer mortality statistics in the united states. *Cancer Epidemiology*, 36(1) :22–28, 2012.

[Fle96] L. Fleury. *Extraction de connaissances dans une base de données pour la gestion des ressources humaines*. PhD thesis, Université de Nantes, 1996.

[Fre98] AA. Freitas. On objective measures of rule surprisingness. In *European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 1–9. Springer, 1998.

[Fre00] AA. Freitas. Understanding the crucial differences between classification and discovery of association rules : a position paper. *AcM sIGKDD Explorations Newsletter*, 2(1) :65–69, 2000.

[GA96] R. Gras and SA. Almouloud. *L'implication statistique : nouvelle méthode exploratoire de données : applications à la didactique*. La pensée sauvage, 1996.

[GCB+04] R. Gras, R. Couturier, J. Blanchard, H. Briand, P. Kuntz, and P. Peter. Quelques critères pour une mesure de qualité de règles d'association. *Revue des nouvelles technologies de l'information RNTI E-1*, pages 3–30, 2004.

[GCG15] R. Gras, R. Couturier, and P. Gregori. Un mariage arrange entre l'implication et la confiance. In *8th International Meeting Statistical Implicative Analysis. Tunisia : Institut Supérieur des Études Technologiques de Radès*, 2015.

[GDRG06] R. Gras, J. David, JC. Régnier, and F. Guillet. Typicalité et contribution des sujets et des variables supplémentaires en analyse statistique implicative. In *EGC*, pages 359–370, 2006.

[GH07] F. Guillet and HJ. Hamilton. *Quality measures in data mining*, volume 43. Springer, 2007.

[Gha22] S. Ghanem. An r package for sia binary classification. https ://github.com/souhilabsl/SIAclassification, accessed on 23/04/2022).

[GKB01] R. Gras, P. Kuntz, and H. Briand. Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données. *Mathématiques et sciences humaines. Mathematics and social sciences*, (154), 2001.

[GKB03] R. Gras, P. Kuntz, and H. Briand. Hiérarchie orientée de règles généralisées en analyse implicative. In *EGC*, pages 145–157, 2003.

[GKCG01] R. Gras, P. Kuntz, R. Couturier, and F. Guillet. Une version entropique de l'intensité d'implication pour les corpus volumineux. *Actes des journées Extraction et Gestion des Connaissances (EGC)*, (187) :69–80,, 2001.

[GKG15] R. Gras, P. Kuntz, and N. Greffard. Notion de champ implicatif en analysis statistique implicative. In *The 8th International Meeting on Statistical Implicative Analysis, Tunisia*, pages 1–21, 2015.

[GR13] R. Gras and JC. Régnier. Analyse implicative des variables binaires. intensité implicative. intensité entropique, 2013.

[GR17] R. Gras and JC. Regnier. Analyse implicative des variables binaires. intensité implicative. intensité entropique (partie 1, ch. 1), 2017.

[Gra79] R. Gras. *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques*. Université de Rennes I France, 1979.

[Gra96] R. Gras. Nouvelle méthode exploratoire de données. *La Pensée sauvage, éditions, Francia*, 1996.

[Gra00] R. Gras. Les fondements de l'analyse implicative statistique. *Quaderni di Ricerca in Didattica*, 2000.

[GRG09] R. Gras, JC. Régnier, and F. Guillet. Analyse statistique implicative. une méthode d'analyse de données pour la recherche de causalités. *Toulouse (França) : Cepadues*, 2009.

[Gro14] US Cancer Statistics Working Group. United states cancer statistics : 1999–2010 incidence and mortality web-based report. *Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute*, 2014.

[GS88] R. Goodman and P. Smyth. Information-theoretic rule induction. In *Proceedings of the 8th European Conference on Artificial Intelligence*, pages 357–362, 1988.

[GSGS08] R. Gras, E. Suzuki, F. Guillet, and F. Spagnolo. *Statistical implicative analysis*. Springer, 2008.

[Gui02] S. Guillaume. Discovery of ordinal association rules. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 322–327. Springer, 2002.

[Har15] F. Harrell. http ://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/CstressEcho.html, 2015.

[HCHB11] M. Hahsler, S. Chelluboina, K. Hornik, and C. Buchta. The arules r-package ecosystem : analyzing interesting patterns from large transaction data sets. *The Journal of Machine Learning Research*, 12 :2021–2025, 2011.

[HFC20] C. Shi H. Fang and CH. Chen. Bioexpdnn : Bioinformatic explainable deep neural network. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2461–2467. IEEE, 2020.

[HGG00] J. Hipp, U. Güntzer, and G.Nakhaeizadeh. Algorithms for association rule mining-a general survey and comparison. *ACM sigkdd explorations newsletter*, 2(1) :58–64, 2000.

[HHNT86] J. Holland, K. Holyoak, R. Nisbett, and P. Thagard. Induction : Processes of inference, learning and discovery. *MIT Press*, 1986.

[HPK11] J. Han, J. Pei, and M. Kamber. *Data mining : concepts and techniques*. Elsevier, 2011.

[IHSM15] AE. Ibrahim, AI. Hashad, NEM. Shawky, and A. Maher. Robust breast cancer diagnosis on four different datasets using multi-classifiers fusion. *Int. J. Eng. Res. Technol.(IJERT)*, 4 :114–118, 2015.

[IUY13] O. Inan, MS. Uzer, and N. Yılmaz. A new hybrid feature selection method based on association rules and pca for detection of breast cancer. *International Journal of Innovative Computing, Information and Control*, 9(2) :727–729, 2013.

[JBA99] J. Roberto Jr. Bayardo and R. Agrawal. Mining the most interesting rules. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 145–154, 1999.

[KBC17] H. Khaled, A. Bodin, and R. Couturier. L'analyse statistique implicative (asi) appliquée aux données de programme international pour le suivi des acquis des élèves (pisa). *Points de vue conceptuels, applicatifs et métaphoriques*, page 437, 2017.

[Kod00] Y. Kodratoff. Extraction de connaissances à partir des données et des textes». *Actes des journées sur la fouille dans les données par la méthode d ?analyse statistique implicative, Presses de l'Université de Rennes*, 1 :151–165, 2000.

[LA02] IC. Lerman and J. Azé. Indice probabiliste discriminant (de vraisemblance du lien) d'une règle d'association en cas de très grosses données. *Contribution au rapport d'activité du Groupe Qualité de l'action GaFo-Données*, 2002.

[Lag98] JB. Lagrange. Analyse implicative d'un ensemble de variables numériques ; application au traitement d'un questionnaire à réponses modales ordonnées. *Revue de statistique appliquée*, 46(1) :71–93, 1998.

[Lal02] S. Lallich. Mesure et validation en extraction des connaissances à partir des données. *Habilitationa Diriger des Recherches–Université Lyon*, 2, 2002.

[Ler] IC. Lerman. Classification et analyse ordinale des données, 1981. *Dunod, Paris.*

[LHM+98a] B. Liu, W. Hsu, Y. Ma, et al. Integrating classification and association rule mining. In *Kdd*, volume 98, pages 80–86, 1998.

[LHM+98b] B. Liu, W. Hsu, Y. Ma, et al. Integrating classification and association rule mining. In *Kdd*, volume 98, pages 80–86, 1998.

[LHP01] W. Li, J. Han, and J. Pei. Cmar : Accurate and efficient classification based on multiple class-association rules. In *Proceedings 2001 IEEE international conference on data mining*, pages 369–376. IEEE, 2001.

[LLV05] St. Lallich, P. Lenca, and B. Vaillant. Variations autour de l'intensité d'implication. In *3d International Conference Implicative Statistic Analysis, October 6-8, Palermo, Italia*, pages 237–246. Gras, R. and Spagnolo, F. and David, J., 2005.

[Loe47] JE. Loevinger. A systematic approach to the construction and evaluation of tests of ability. *Psychological monographs*, 61(4) :i, 1947.

[LR98] D. Lahanier-Reuter. *Étude de conceptions du hasard : approche épistémologique, didactique et expérimentale en milieu universitaire.* PhD thesis, Rennes 1, 1998.

[LT04] S. Lallich and O. Teytaud. Évaluation et validation de l'intérêt des regles d'association. *Revue des nouvelles Technologies de l'Information*, 1(2) :193–218, 2004.

[LVL07] St. Lallich, B. Vaillant, and P. Lenca. A probabilistic framework towards the parameterization of association rule interestingness measures. *Methodology and Computing in Applied Probability*, 9(3) :447–463, 2007.

[Mal10] St. Malaise. Classification hiérarchique de compétences par l'intermédiaire du logiciel chic fondé sur la méthode d'analyse statistique implicative. *Actes du congrès international d ?Actualités de la Recherche en Éducation et en Formation AREF 2010*, 2010.

[MC11] St. Malaise and D. Casanova. Hiararchisation de competences par l'intermediaire du logiciel chic : Exemples d'analyses de données dans deux domains différent. In *24e Colloque International de L'ADMEE, 2011-08-20*, 2011.

[MF06] H. Mohamadally and B. Fomani. Svm : Machines a vecteurs de support ou separateurs a vastes marges. *Survey, Versailles St Quentin*, 16, 2006.

[MMH97] T. Mitchell and M. McGraw-Hill. Edition, 1997.

[MRD14] PL. Micheaux and B. Liquet R. Drouilhet. Présentation du logiciel r. In *Le logiciel R*, pages 1–29. Springer, 2014.

[MW+16] R. Machmud, A. Wijaya, et al. Behavior determinant based cervical cancer early detection with machine learning algorithm. *Advanced Science Letters*, 22(10) :3120–3123, 2016.

[O+96] MJL. Orr et al. Introduction to radial basis function networks, 1996.

[OGC05] P. Orús, P. Gregori, and Campus Riu Sec-ESTCE Castellón. Des variables supplémentaires et des élèves «fictifs», dans la fouille didactique de données avec chic. *Troisième rencontre internationale de l'Analyse Statistique Implicative (ASI3)*, pages 279–291, 2005.

[Par04] M. Parizeau. Réseaux de neurones. *GIF-21140 et GIF-64326*, 124, 2004.

[Pea96] K. Pearson. Vii. mathematical contributions to the theory of evolution.iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, (187) :253–318, 1896.

[Pla98] J. Platt. Sequential minimal optimization : A fast algorithm for training support vector machines. 1998.

[PS91] G. Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. *Knowledge discovery in databases*, pages 229–238, 1991.

[PV98] M. Pontil and A. Verri. Support vector machines for 3d object recognition. *IEEE transactions on pattern analysis and machine intelligence*, 20(6) :637–646, 1998.

[Q+92] JR. Quinlan et al. Learning with continuous classes. In *5th Australian joint conference on artificial intelligence*, volume 92, pages 343–348. World Scientific, 1992.

[Qui92] JR. Quinlan. Learning with continuous classes v : 5th australian joint conference on artificial intelligence. adams a., sterling l.(eds.). hobart, tasmania, 1992.

[Qui93] RC. Quinlan. 4.5 : Programs for machine learning morgan kaufmann publishers inc. *San Francisco, USA*, 1993.

[Ram08] G. Ramstein. Statistical implicative analysis of dna microarrays. In *Statistical Implicative Analysis*, pages 205–225. Springer, 2008.

[Ren80] T. René. Paraboles et catastrophes.[1983 paris : Flammarion.]. 1980.

[SAZ12] GI. Salama, M. Abdelhalim, and MA. Zeid. Breast cancer diagnosis on three different datasets using multi-classifiers. *Breast Cancer (WDBC)*, 32(569) :2, 2012.

[SC95] D. Steinberg and P. Colla. Cart : tree-structured non-parametric data analysis. *San Diego, CA : Salford Systems*, 1995.

[SC97] D. Steinberg and P. Colla. Cart : Classification and regression trees ; salford systems. *San Diego, CA., USA*, 1997.

[SGK00] A. Kenchaf S. Guilaume and P. Kuntz. Ordinal intensity of implication for discretization. *Conf. on Ordinal and Symbolic Data Analysis, Bruxelles*, 2000.

[Sha01] CE. Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1) :3–55, 2001.

[SK13] PK. Srimani and MS. Koti. Medical diagnosis using ensemble classifiers-a novel machine-learning approach. *Journal of Advanced Computing*, 1(6) :9–27, 2013.

[SMK] K. SABANCI and year=2015 M. KOKLU. The classification of eye state by using knn and mlp classification models according to the eeg signals.

[SS88] M. Sebag and M. Schoenauer. Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases. In *Proc. of EKAW*, volume 88, page 28, 1988.

[ST09] M. Steinbach and P.N. Tan. knn : k-nearest neighbors. *The top ten algorithms in data mining*, pages 151–162, 2009.

[S.W83] S.Wright. On" path analysis in genetic epidemiology : a critique". *American journal of human genetics*, 35(4) :757, 1983.

[TCP04] F. Thabtah, P. Cowling, and Y. Peng. Mmac : A new multi-class, multi-label associative classification approach. In *Fourth IEEE International Conference on Data Mining (ICDM'04)*, pages 217–224. IEEE, 2004.

[TF+19] A. Totohasina, DR. Feno, et al. An extension of totohasina's normalization theory of quality measures of association rules. *International Journal of Mathematics and Mathematical Sciences*, 2019, 2019.

[TKS04] PN. Tan, V. Kumar, and J. Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4) :293–313, 2004.

[TM18] H. Taud and JF. Mas. Multilayer perceptron (mlp). In *Geomatic approaches for modeling land change scenarios*, pages 451–455. Springer, 2018.

[TW14] KL. Man T. Wang, SU. Guan. Time series classification for eeg eye state identification based on incremental attribute learning. In *2014 International Symposium on Computer, Consumer and Control*, pages 158–161. IEEE, 2014.

[VL08] B. Vo and B. Le. A novel classification algorithm based on association rules mining. In *Pacific Rim Knowledge Acquisition Workshop*, pages 61–75. Springer, 2008.

[VLL04] B. Vaillant, P. Lenca, and S. Lallich. Etude expérimentale de mesures de qualité de règles d'associations. In *EGC 2004 : Extraction et Gestion des Connaissances, 20-23 janvier, Clermont-Ferrand, France*, pages 341–352. Cépadues, 2004.

[VLL08] B. Vaillant, St. Lallich, and P. Lenca. On the behavior of the generalizations of the intensity of implication : A data-driven comparative study. In *Statistical Implicative Analysis*, pages 421–447. Springer, 2008.

[WFM92] G. Piatetsky-Shapiro WJ. Frawley and CJ. Matheus. Knowledge discovery in databases : An overview. *AI magazine*, 13(3) :57–57, 1992.

[YH03] X. Yin and J. Han. Cpar : Classification based on predictive association rules. In *Proceedings of the 2003 SIAM international conference on data mining*, pages 331–335. SIAM, 2003.

[YYH+10] Z. You, Z. Yin, K. Han, D. Huang, and X. Zhou. A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network. *Bmc Bioinformatics*, 11(1) :1–13, 2010.

Résumé

Dans cette thèse nous nous sommes intéressés à une approche statistique originale dédiée à l'extraction et à l'analyse des règles. Il s'agit de l'Analyse statistique implicative (ASI) qui a la particularité d'être une méthode non symétrique basée sur la mesure de l'intensité d'implication. Elle mesure l'étonnement d'observer un petit nombre de contre-exemples à une règle. Notre premier objectif dans cette thèse est de montrer l'apport de l'ASI et l'efficacité du logiciel (R)CHIC qui implémente les fonctionnalités de l'ASI. Cet objectif a été réalisé à travers un état de l'art sur l'ASI et une étude des notes des étudiants en informatique de l'université de Bejaia. Le deuxième objectif consiste à améliorer et créer certaines fonctions permettant de calculer les méthodes de l'ASI dans le logiciel R et rajouter d'autres fonctionnalités au logiciel RCHIC que CHIC ne possède pas. Pour cela un nouveau mode de calcul dans le graphe implicatif a été conçu, il offre la possibilité à l'utilisateur d'avoir des règles étonnantes auquel on peut faire confiance. Le troisième objectif a été de concevoir deux classificateurs ayant du sens et facile à interpréter avec les règles d'implication de l'ASI.

Mots clés : ASI, (R)CHIC, Graphe Implicatif, Classification.

Abstract

In this thesis we are interested in an original statistical approach dedicated to the extraction and analysis of rules. That is the Statistical Implicative Analysis (SIA) which has the particularity of being a non-symmetrical method based on the intensity of implication measurement. It measures the surprise of observing a small number of counter-examples to a rule. Our first objective in this thesis is to illustrate the contribution of SIA and the efficiency of the (R)CHIC software, that implements the SIA functionalities. This objective was achieved through a state of the art on SIA and by studying the grades of computer science students at the Bejaia University. The second goal is to improve and create some functions to compute ASI methods in the R software and add other features to RCHIC that CHIC does not have. In order to do this, a new mode of calculation in the implicative graph has been designed, it offers the user the possibility to have surprising rules that can be trusted. The third objective was to design two classifiers that make sense and easy to interpret with the SIA implication rules..**Key words:** SIA, (R)CHIC, Implicatif graph, Classification

الملخص

في هذه الأطروحة، نحن مهتمون بطريقة إحصائية أصلية مخصصة لاستخراج القواعد وتحليلها والتي تسمى طريقة تحليل المعطيات (ASI) التي تتميز بكونها طريقة غير متناظرة ترتكز على مقياس حدة الاشتراك بين عناصر الدراسة والتي تثمن الهدف من قاعدة معينة بالنظر إلى عدد الأمثلة المضادة. هدفنا الأول في هذه الأطروحة هو إظهار مدى أهمية هذه الطريقة ومدى كفاءة برنامج (RCHIC) الذي يفعل وظائف طريقة (ASI). تم تحقيق هذا الهدف من خلال دراسة تقنيات (ASI) ودراسة عينة من نتائج شريحة طلبة الإعلام الآلي على مستوى جامعة بجاية. يتمثل الهدف الثاني في تحسين وإنشاء وظائف معينة مما يجعل من الممكن حساب طرق (ASI) في برنامج (R) وإضافة وظائف أخرى إلى برنامج (RCHIC) التي لا تتوفر لدى (CHIC). لهذا تم تصميم وضع جديد للحساب في الرسم البياني الضمني، وهو يوفر إمكانية للمستخدم أن يكون لديه قواعد مذهلة يمكن الوثوق بها. الهدف الثالث يتمثل في تصميم مصنفين لهما معنى ويسهل تفسيرهما باستخدام قواعد تضمين ASI.

الكلمات المفتاحية: ASI ،(R)CHIC ،الرسم البياني الضمني ،تصنيف.