République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université ABDERRAHMANE MIRA de Bejaïa
Faculté des Sciences Exactes
Département D'Informatique

Mémoire de fin de Cycle
En vue de l'obtention du diplôme de Master en informatique
Option : Administration et sécurité réseaux

Thème

Application de l'Analyse Statistique Implication (L'ASI) dans le domaine de la santé

Réalisé par :

Mr MEHIDI Walid-fatsah

Mr MEHDIOUI Ali

Devant le Jury composé de :

Présidente	Mme	Houda	EL BOUHISSI	M.C.B	Université de Bejaia
Examinatrice	Mme	Hayette	KHALED	M.A.A	Université de Bejaia
Promotrice	Mme	Souhila	GHANEM	M.A.A	Université de Bejaia

Année Universitaire : 2019/2020

Remerciement

En premier lieu, nous tenons à remercier le bon Dieu de nous avoir donné la santé, le courage et la volonté durant la préparation de ce modeste travail.

Nous tenons à remercier vivement Mme GHANEM Souhila, pour nous avoir honorés par son encadrement, pour sa disponibilité, ses orientations, ses précieux conseils et ses encouragements qui nous ont permis de mener à bien ce travail.

Nous poursuivons ces remerciements en saluant vivement les membres du jury pour l'honneur qu'ils nous ont fait en acceptant de juger ce travail.

Nous n'omettrons jamais d'exprimer toute notre gratitude à tous les membres du département d'informatique de l'université de Bejaïa, que ce soit enseignants ou cadres administratifs, qui de près ou de loin n'ont épargné aucun effort pour que notre formation et nos travaux se termine dans de bonnes conditions.

Un merci particulier à nos parents, pour leur amour, leurs sacrifices et leurs patiences.

Un énorme merci à nos familles et amis pour leurs éternel soutient et la confiance qu'ils ont en nos capacité.

Dédicaces

Ce modeste travail est dédié :

A nos chers parents qui nous ont soutenus et encouragés durant toute

Notre scolarité.

A nos frères et sœurs

A nos enseignants

A nos amis(e)

A toutes les personnes qui nous ont apportés de l'aide.

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

ASI	Analyse statistique implicative
AFC	l'Analyse Factorielle des Correspondances
CAH	la Classification Ascendante Hiérarchique
CHIC	Acronyme de Classification Hiérarchique Implicative et Cohésitive
WBC	Wisconsin Breast Cancer
WDBC	Wisconsin Diagnosis Breast Cancer

Introduction générale

L'analyse ou la fouille dans les données part en général, du croisement de sujets (ou objets) et de variables (propriétés ou attributs) binaires, ordinales ou numériques. Son objectif majeur consiste à conjecturer des modèles basés sur des relations quantitatives ou qualitatives et des structures induites à partir des données. Différentes méthodes, comme l'Analyse Factorielle des Correspondances (A.F.C.), la Classification Ascendante Hiérarchique (C.A.H.), sont communément utilisées pour de telles fouilles dans des données. Parmi elles, l'Analyse Statistique Implicative (A.S.I.), née de problématiques didactiques en mathématiques, fondée et développée par Régis Gras et son équipe vise l'extraction de connaissances, d'invariants, de règles inductives non symétriques consistantes.

Nouvelle en tant que méthode multidimensionnelle d'analyse non symétrique de données, l'Analyse Statistique Implicative (A.S.I.), croise un ensemble de sujets ou d'objets et un ensemble de variables. Elle complète, voire remplace des méthodes docimologiques, corrélationnelles ou/et psychométriques traditionnelles. L'A.S.I par ses extensions diverses, se présente maintenant comme une large méthode d'Intelligence Artificielle visant l'extraction de causalités sous forme de règles, dans un ensemble de variables de nature variée. Elle est basée, de façon originale, sur l'invraisemblance de l'existence de ces relations, c'est-à-dire sur la faiblesse relative de leurs contre-exemples par rapport à ce que donnerait le hasard seul. Elle établit une relation topologique duale entre l'ensemble de sujets et celui des variables. De nombreuses applications de cette approche, moteurs ou creusets de développement de l'A.S.I., ont concerné et concernent encore des domaines variés comme la psychologie, la sociologie, la médecine, la biologie, l'économie, l'histoire de l'art, etc. Elles font l'objet d'autres ouvrages sur l'A.S.I

Notre but dans ce mémoire est :

- Premièrement, faire reconnaître l'ASI a la communauté universitaire ici à Bejaia, vu que c'est un nouveau domaine très peut utiliser ici en Algérie mais qui peut apporter beaucoup d'amélioration dans le domaine d'éducation, de la médecine,….etc.

- Comparer entre les deux options offertes par le logiciel qui traite L'ASI (RCHIC) (intensité d'implication associé a la confiance et implifiance) Pour cela nous avons utilisé des données sur le cancer de sein
- Détecter les causes du cancer de sein.

Pour mener à bien notre travail nous l'avons organisé en trois chapitres comme suit :

Chapitre 1 : Généralités sur l'analyse statistique implicative : Nous présentons les différents concepts de cette méthode tels que les spécificités, représentations graphiques et domaines d'applications.

Chapitre 2 : L'environnement de travail : Nous allons présenter notre environnement de travail à savoir le logiciel R, Rstudio et Rchic ainsi les étapes d'installation de ces derniers.

Chapitre 3 : Application de l'analyse statistique implicative : ou nous présenterons les résultats appliqués aux données grâce au graphe implicatif

En fin de ce mémoire, une conclusion est donnée pour résumer les apports essentiels de notre Travail.

Chapitre 1

Généralités sur l'analyse statistique implicative

1.1. Introduction :

L'Analyse Statistique Implicative (A.S.I.), croise un ensemble de sujets ou d'objets et un ensemble de variables. Elle complète, voire remplace des méthodes docimologiques, corrélationnelles ou/et psychométriques traditionnelles, aujourd'hui elle concerne des domaines variés comme la psychologie, la sociologie, la médecine, la biologie, l'économie, bio-informatique, l'histoire de l'art, etc.

Dans ce chapitre nous allons expliquer le concept de l'ASI, en présentant un bref historique détaillée, ses spécificités ainsi la démarche mathématiques appliquer et quelques exemples d'utilisation de cette méthode.

1.2. Historique :

Au cours des années 70, dans le cadre des Instituts de Recherche sur l'enseignement des Mathématique de France Régis Gras a fréquenté des classes du secondaire particulièrement du 1 er cycle (11 à 15 ans) ou il a conduit et évalué une expérience nationale, il disposait de données constituées de traces laissées par les élèves. A l'occasion de résolution d'exercices de mathématiques ou de problèmes, une certaine hiérarchie de difficultés segmentait l'ensemble des élèves interrogés. Plus la difficulté s'accroissait plus le

nombre de réussites diminuait ce qui peut sembler une tautologie : il peut en effet s'attendre à ce que tout élève qui réussit une épreuve jugée difficile, dans un contexte qui serait comparable réussirait a fortiori ce qui était facile. Ce qui pourrait étonner et le contester ce sont les incohérences par rapport à cet attendu ce qui signifie que des élèves pouvaient dans certains cas et pour quelques-uns d'entre eux réussir à un item *a* jugé difficile tout en échouant à un item b jugé plus facile et ceci sans remettre en cause l'affirmation que «généralement la réussite à *a* s'accompagne de la réussite à *b* » et sans que sa réciproque ne soit nécessairement vraie, son intérêt va alors porter sur ce type de relation non symétrique..

Tous les outils statistiques qui étaient à sa disposition pour qualifier et quantifier cette relation étaient incertains, la stratégie qu'il a alors utilisée, en 1978, a consisté à prendre plutôt en considération la non-satisfaction de l'implication « si *a* alors *b* » qui, comme on le sait, apparaît dès que a étant vrai, b est faux Ce sont donc les contre-exemples sur lesquels va porter son attention, ce qui fait que le raisonnement suivis est asymétrique. [B1]

1.3. Définition de l'ASI :

L'analyse statistique implicative est une méthode non symétrique d'analyse de données créée par Régis Gras et qui a un impact significatif sur divers domaines allant de la recherche pédagogique et psychologique à l'exploration de données. D'une part permet de déceler les règles pertinentes à partir d'un test d'hypothèse sur des données variées, d'autre part offre selon une démarche calquée sur la classification hiérarchique classique, une représentation hiérarchique des méta-règles et une analyse des contributions des attributs et individus aux différentes associations.

L'implication statistique au sens de Gras, s'exprime en termes de règle de la quasi-implication a implique statistiquement b si b a tendance à être vérifier quand a l'est, la force ou la qualité de la règle s'apprécie à l'aide d'un indice (implication statistique) qui mesure l'invraisemblance du nombre de contre-exemples à la règle où la règle n'est pas vérifiée.

L'ASI c'est une théorie générale dans le domaine de la causalité parce qu'elle répond à des faiblesses d'autres théories de faite de prendre en compte la contraposée de la règle parce qu'on s'intéresse à des règles causales et non descriptives, la prise en compte concomitante de la contraposée de l'implication est indispensable pour renforcer l'évaluation de la qualité suffisamment bonne de la relation de quasi-implication, voire quasi-causale, de a sur b. En

même temps, elle pourrait permettre de corriger la difficulté évoquée en relation à la taille des ensembles en jeu. [B1]

1.4. Les spécificités de l'analyse statistique implicatives :

L'ASI, mesurée à d'autres méthodes d'analyse de données, présente de caractères originaux importants on les résume :

- Modèles successifs de variables répondant à des contraintes épistémologiques explicites compatibles avec la sémantique des situations à modéliser
- Non-symétrie de la méthode
- Capacités pédagogiques et ergonomiques des représentations en particulier pour l'examen des règles généralisées
- Dualité structurelle des deux espaces en jeu : sujet et variables avec les notions de contributions et typicalités aux structures
- La simplicité du modèle mathématique sous-jacent lui assurant accessibilité plasticité et fécondité utiles pour répondre a des attentes applicatives dans de larges domaines
- Extension du discret fini au continu tant pour les variables que pour les sujets.
- Extension progressive de la nature des variables traitées tout en conservant les propriétés de plongement. [B3]

1.5. Démarche mathématique :

Pour ce faire, à l'instar de la méthode de mesure de la similarité de I.C. Lerman (1981), à l'instar de la démarche classique dans les tests non paramétriques (ex. Fischer, Wilcoxon, etc.), Gras 1979, Gras et al. 1996 définissons la mesure de qualité confirmatoire de la relation implicative a=>b à partir de l'invraisemblance de l'apparition, dans les données, du nombre (estimateur) de cas qui l'infirment, c'est-à-dire pour lesquels a est vérifié sans que b ne le soit. Ceci revient à comparer l'écart entre le contingent et le théorique si seul le hasard intervenait. Mais, dans le cadre de l'analyse de données, c'est cet écart qui est pris en compte et non pas l'énoncé d'un rejet ou de l'admissibilité d'hypothèse nulle. Cette mesure est relativisée par le nombre de données vérifiant respectivement a et non b, circonstance dans laquelle l'implication est précisément mise en défaut. Elle quantifie "l'étonnement" de l'expert

devant le nombre invraisemblablement petit de contre-exemples sous l'hypothèse d'une indépendance entre les variables eu égard aux effectifs en jeu.

Un ensemble fini V de v variables, désignées par des lettres a, b, c, etc., est donné. Dans la situation paradigmatique classique, il s'agit des performances (réussite-échec) à des items d'un questionnaire de connaissances. A un ensemble fini E de n sujets désignés x, on associe, par abus d'écriture, les fonctions du type : x=>a(x) où a(x)=1 (ou a(x)=vrai) si x satisfait ou possède le caractère a et 0 (ou a(x)=faux) sinon. En intelligence artificielle, on dira que x est un exemple ou une instance pour a si a(x)=1 et un contre-exemple dans le cas contraire.[B2]

La règle a=>b est logiquement vraie si pour tout x de l'ensemble E, b(x) n'est nul que dans le cas où a(x) l'est aussi, autrement dit si l'ensemble A des x pour lesquels a(x)=1 est contenu dans l'ensemble B des x pour lesquels b(x)=1. Cependant, cette inclusion stricte n'est qu'exceptionnellement observée dans les expériences réelles. Dans le cas d'un questionnaire de connaissances, on pourrait en effet observer quelques rares élèves réussissant un item a et ne réussissant pas l'item b, sans que ne soit contestée la tendance à réussir b quand on a réussi a. Relativement aux cardinaux de E (soit n), de A (soit na) et de B (soit nb), c'est le poids des contre-exemples (soit $n_{a\wedge\bar{b}}$) qu'il faut donc prendre en compte pour accepter statistiquement de conserver ou non la quasi-implication ou quasi-règle a=>b. Ainsi, c'est à partir de la dialectique entre les exemples et les contre-exemples que la règle apparaît comme le dépassement de la contradiction. [B2]

1.5.1. Formalisation de la quasi-règle implicative dans l'approche classique :

Pour formaliser cette quasi-règle, ils considèrent, comme le fait I.C. Lerman pour la similarité, 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 ensembles complémentaires respectifs de Y et de B dans E de même cardinal $n_{\bar{b}} = n\text{-}n_b$

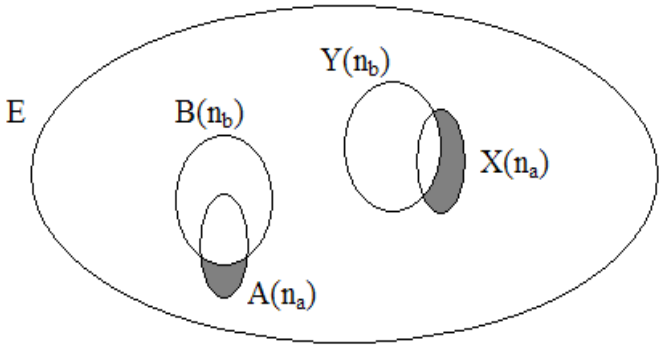

Figure 1.1: représentation ensembliste

Soit α un réel quelconque de l'intervalle [0,1]

La quasi-règle a=>b est admissible au niveau de confiance 1-α si et seulement si $\Pr\left[Card(X \cap \overline{Y}) \leq card(A \cap \overline{B})\right] \leq \alpha$

Intuitivement et qualitativement, ceci signifie que la quasi-implication a=>b sera admissible à l'issue d'une expérience si le nombre d'individus de E la contredisant est invraisemblablement petit par rapport au nombre d'individus attendu sous une hypothèse d'absence de lien. Par exemple, si Card E = 100, Card A = 36, Card B = 50, alors $card(A \cap \overline{B}) = 3$ est «invraisemblablement petit» sous l'hypothèse d'une absence de lien entre a et b. ils constatent en effet que A est "presque" contenu dans B, alors que, sans liaison de A et B, on pourrait s'attendre à ce qu'environ la moitié des éléments de A soient aussi dans B. [B3]

1.5.2. Intensité d'implication :

On appelle intensité d'implication de la quasi-règle *a=>b* le nombre :

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

L'intensité d'implication est une valeur probabiliste, et non une fréquence, qui fonde la décision de retenir ou non une relation de quasi-implication entre les variables binaires a et b. Cette modélisation de la quasi-implication est pertinente pour mesurer l'étonnement face au constat de la petitesse du nombre des contre-exemples en regard du nombre surprenant des instances de l'implication. Il s'agit d'une mesure de la qualité inductive et informative de l'implication. Par conséquent, si la règle est triviale, comme dans le cas où B est très grand ou coïncide avec E, cet étonnement devient petit. Gras R., (1996) a démontré d'ailleurs que cette trivialité se traduit par une intensité d'implication très faible, voire nulle : Si, na étant fixé et A

étant inclus dans B, nb tend vers n (B "croît" vers E), alors ϕ(a,b) tend vers 0. C'est pourquoi ils définit par « continuité »: ϕ(a,b) = 0 si nb = n. De même, si A⊂ B, ϕ (a,b) peut être inférieure à 1 dans le cas où la confiance inductive, mesurée par l'étonnement statistique, est insuffisante.[B3]

1.5.3. Intensité entropique :

Deux raisons qui ont conduits à améliorer le modèle formalisé par l'intensité d'implication dans l'approche classique :

lorsque les tailles des ensembles d'individus traités augmentent, atteignant des effectifs de l'ordre du millier ou plus l'intensité d'implication ϕ(a,b) a tendance à ne plus être suffisamment discriminante car ses valeurs peuvent être très voisines de 1 alors que l'inclusion dont elle cherche à modéliser la qualité, est loin d'être satisfaite. Ce phénomène a été déjà signalé par A. Bodin [B11], dont les travaux traitent avec des ensembles de grande taille d'élèves impliqués dans des enquêtes internationales.

Le modèle classique de la quasi-implication retient essentiellement la mesure de l'intensité de la quasi-règle a=>b, *la* prise en compte concomitante de la contraposée de l'implication de non b=>non *a* est indispensable pour renforcer l'évaluation de la qualité suffisamment bonne de la relation de quasi-implication, voire quasi-causale, de a sur b. En même temps, elle pourrait permettre de corriger la difficulté évoquée en relation à la taille des ensembles en jeu. En effet si A et B sont des ensembles de petite taille par rapport à E, leurs complémentaires seront importants et réciproquement.

La solution utilise à la fois l'intensité d'implication et un autre indice qui rend compte de la dissymétrie entre les situations S1= (a et b) et S'1= (a et non b) qui concerne la quasi-règle a=>b ainsi que celle entre les situations S2= (non a et non b) et S'2 = (a et non b) qui concerne la quasi-règle contraposée. [B2]

1.5.4. L'implifiance :

On appelle implifiance la mesure de l'implication statistique qui prend en compte l'implication directe et sa contraposée ainsi que la confiance. [B3]

1.6. L'intensité d'implication et d'autres indices :

Nous allons présenter d'autres indices qui existent et donner les propriétés d'un indice de qualité et faire une comparaison entre l'intensité d'implication (indice utilisé avec l'ASI) et d'autres indices.

1.6.1. Définition de quelques indices :

1.6.1.1. La confiance : Elle indique la proportion d'entités vérifiant le conséquent parmi celles vérifiant la prémisse de la règle. Cette mesure est non symétrique et non Implicative. Elle prend la valeur de référence 1/2 à l'équilibre. Elle n'est pas sensible à la taille de données : c'est donc une mesure descriptive. Elle prend ses valeurs sur l'intervalle [0; 1]. [B3]

1.6.1.2. Le Lift : Cette mesure de qualité représente le rapport d'indépendance entre la prémisse et le conséquent de la règle. Lift est une mesure symétrique non implicative. Il est sensible a la taille de données : c'est une mesure statistique. Lift prend ses valeurs sur [0;+1[. [B10]

1.6.1.3. Coefficient de corrélation : le coefficient de corrélation permet de donner une mesure synthétique de l'intensité de la relation entre deux caractères, il permet d'analyser les relations linéaires et rend compte d'une liaison linéaire entre les variables a et b. [B10]

1.6.1.4. L'indice de Loevinger : C'est un « ancêtre » des indices d'implication (Loevinger, 1947). Cet indice, noté H (a,b) varie de 1 à - ∞. Il est défini par : [B10]

$$H(a,b) = 1 - \frac{n.n_{a \wedge \bar{b}}}{n_a n_{\bar{b}}}$$

1.6.2. Quelques propriétés d'un indice de qualité : [B10]

- ✓ Tenir compte du nombre de contre-exemples
- ✓ Enrichir l'accès à la connaissance par représentation graphique des règles : pour permettre a l'utilisateur d'extraire les relations les plus intéressantes
- ✓ Accepter la variété de types de variables
- ✓ Permettre l'adéquation de la mesure avec la contraposée de la règle : afin de renforcer le caractère quasi-implicatif de a vers b la mesure de a=>b devrait être couplée à celle de sa contraposée non b=>non a, leurs valeurs associées pouvant être très différentes. L'intérêt de cette propriété est pour assurer à la règle une propriété prédictive En effet, si l'implication directe a=>b relie la cause ou les causes conjointes à l'effet, la contraposée non b=>non a nous indique que la disparition de l'effet s'accompagne de

l'extinction des causes, ce qui est manifestement et complémentairement informatif et sémantiquement satisfaisant dans une visée causale.

- ✓ Définir des algorithmes simples et intelligibles : leur complexité ne doit pas conduire à des temps d'exécution trop longs, les algorithmes doivent être programmables de telle façon que l'utilisateur ait accès aux résultats, puisse agir sur les seuils et se donne ainsi le moyen de pratiquer les analyses lui-même.

1.6.3. Comparaison entre l'intensité d'implication et les autres indices : [B3]

- ✓ La variation de l'indice lift est indépendante de celle du nombre de contre exemples, cette indice décroît lorsque le nombre de contre-exemples croît mais la vitesse de décroissance ne dépend pas de la vitesse de croissance de $n_{a\wedge\bar{b}}$.ce qui pareil pour l'intensité d'implication qui diminue lorsque le nombre de contre exemples de l'implication augmente.
- ✓ L'indice de confiance croît quand $n_{a\wedge\bar{b}}$ décroît et la vitesse de variation est constante, indépendante de la vitesse de décroissance de cette quantité ainsi que des variations de n et de n_b
- ✓ Le coefficient de corrélation est symétrique par contre l'indice utilisé dans l'ASI (intensité d'implication) est non symétrique
- ✓ Comme l'intensité d'implication, l'indice de Loevinger est toujours décroissant avec l'augmentation de nombre de contre exemples $n_{a\wedge\bar{b}}$

1.7. Représentations graphiques associées :

Des représentations graphiques permettent de donner toutes les implications étonnantes et de visualiser les relations extraites parmi les graphes nous trouverons :

1.7.1. Graphe implicatif :

Le calcul du graphe implicatif permet d'obtenir un graphe sur lequel les variables qui possèdent une Intensité d'implication supérieure à un certain seuil sont reliées par une flèche représentant l'implication. La figure 2.1 représente un exemple de graphe implicatif. CHIC permet de sélectionner 4 seuils différents et modulables d'implication. L'utilisateur

peut disposer les valeurs comme il le souhaite. Les différentes options permettent par exemple de faire apparaitre les fermetures transitives, de minimiser le nombre de croisements (par l'utilisation d'un algorithme de dessin de graphe).

Il est possible de choisir une zone de travail par défaut et la faire évoluer au fil de l'utilisation. Au début d'un traitement de grande taille, il est préférable de faire intervenir toutes les données et donc de disposer d'une grande surface de travail qui peut être largement supérieure à la taille de l'écran. Puis au cours de l'interprétation, l'utilisateur peut se rendre compte que seules certaines variables lui semblent utiles pour son interprétation. Dans ce cas il supprime temporairement les variables désirées grâce à une boite de dialogue prévu à cet effet. Ensuite, CHIC met à jour à nouveau le graphe des implications. A tout moment il est possible d'ajouter ou de supprimer des variables dans l'analyse que l'on effectue. [B4]

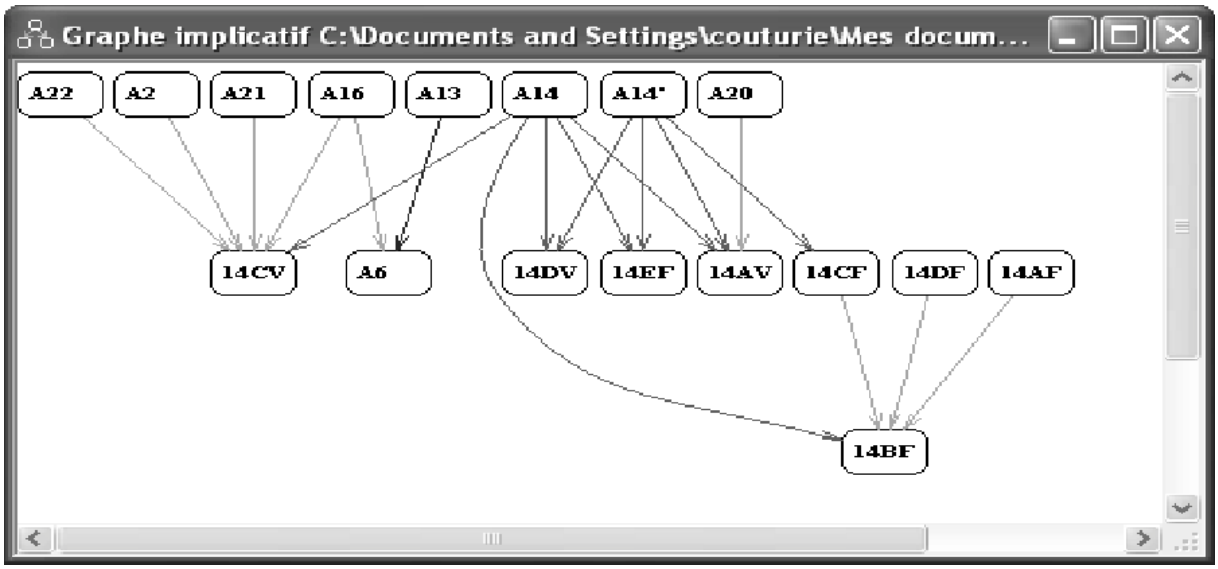

Figure 1.2 :graphe d'implication

1.7.2. L'arbre des similarités :

L'arbre des similarités, figure 1.3, calcule pour chaque couple de variables la similarité entre celles-ci. Ensuite, il agrège des classes constituées elles-mêmes d'autres classes.

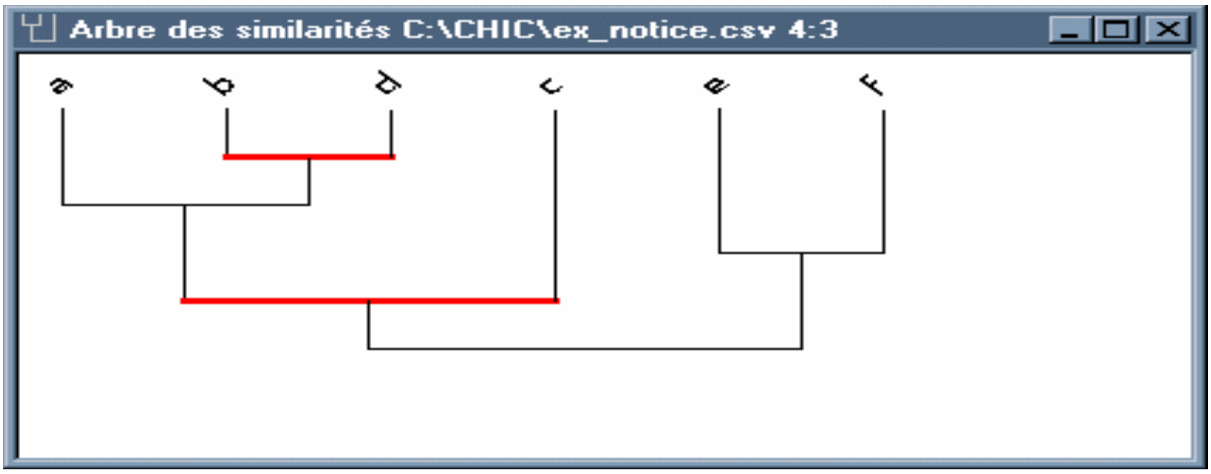

Figure 1.3 : arbre des similarités

Les niveaux identifies par un trait rouge sont des niveaux significatifs dans la mesure ou ceux-ci ont plus de signification classifiant que les autres niveaux. L'algorithme utilisé est l'algorithme de la vraisemblance du lien (AVL) de Lerman (1981). [B4] , sur l'arbre ci-dessus, les variables *b* et *d* sont dans un premier temps les variables les plus similaires. Ensuite l'algorithme choisit d'associer la variables *a* à la classe (*b*,*d*), ainsi la nouvelle classe est (*a*,(*b*,*d*)). A l'itération (ou au niveau) 3, la classe (*e*,*f*) est formée. Finalement au dernier niveau, nous obtenons une seule classe, mais les classes (a,b,d,c) et (e,f) sont dissemblables.

1.7.3. La hiérarchie cohésitive :

L'arbre cohésitif est en première approche, à l'implication ce que l'arbre des similarités est à la similarité. Dans cet arbre, des classes de variables ou de règles entre variables sont constituées à partir des implications entre celles-ci. L'algorithme agrège à chaque étape les variables conduisant à la cohésion la plus forte à cette étape, la figure 1.4 représente un exemple d'un arbre cohesitif. Au premier niveau de la hiérarchie, on remarque que la classe (*b*, *c*).Elle représente le fait que la variable *b* implique la variable *c* avec une intensité plus forte que tous les autres couples de variables. Ce premier niveau de la hiérarchie est d'ailleurs significatif comme l'indique la flèche rouge (en gras sur la figure). Au second niveau, la classe (*a*, (*b*, *c*)) est formée. Cette classe à trois composantes admet la plus forte cohésion parmi celles de toutes les classes possibles à trois composantes et celle de toute autre classe à deux composantes. Puis finalement la classe (*e*, *f*) est créée à la dernière étape. Contrairement à l'algorithme de l'arbre de similarité, l'algorithme construisant l'arbre

cohésitif constitue de manière quasi systématique plusieurs classes et arrête son processus de construction dès que la cohésion entre variables ou entre règles devient trop faible. [B4]

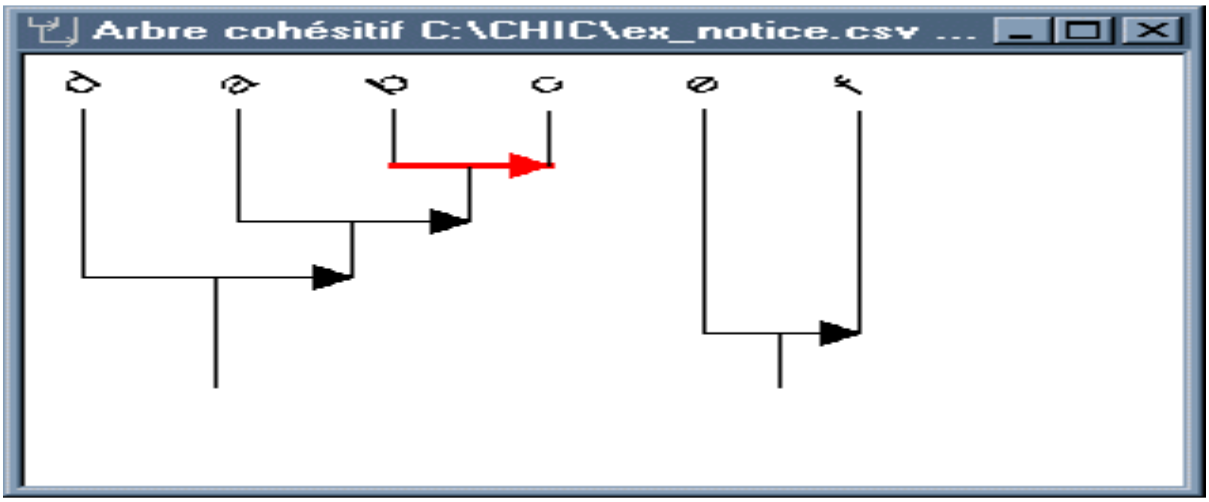

Figure 1.4 : arbre cohesitif

1.7.4. Graphe inclusif :

Le graphe inclusif hérite de la plupart des caractéristiques du graphe implicatif. La différence fondamentale se situe au niveau du calcul de l'inclusion définie par A. BODIN. Plusieurs seuils représentés par plusieurs couleurs permettent de visualiser les différents niveaux d'inclusion. Le changement des valeurs des seuils nécessite de recalculer la totalité des inclusions car les calculs avec d'autres seuils de confiance sont différents. De ce point de vue nous perdons la rapidité offerte par le graphe implicatif lorsque les graphes sont complexes. [B4]

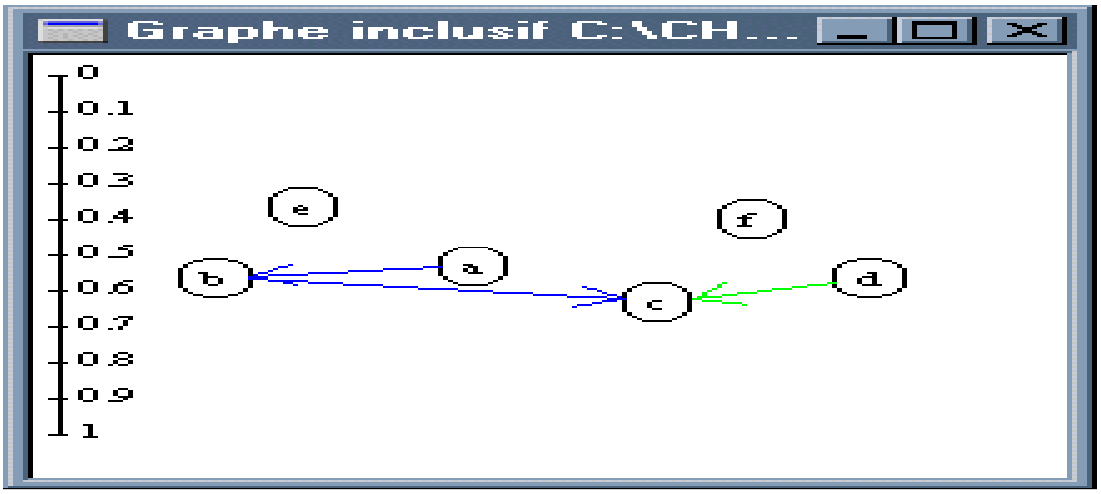

Figure 1.5 : arbre inclusif

1.8. Domaine d'application d'analyse statistique implicative :

De nombreuses applications de développement de l'A.S.I ont concerné et concernent encore des domaines variés, nous allons citer quelques-unes :

1.8.1. Analyse de la dynamique socioprofessionnelle à Genève entre 1816 et 1843 : [B5]

Une expérience d'analyse de statistique implicative afin de détaillée et complétée l'analyse réalisée dans le cadre d'une étude sur le recrutement et le renouvellement des groupes socioprofessionnels à Genève (Oris et al. 2006). Les données considérées résultent de l'appariement deux à deux de 6 recensements. Plus précisément, ils considèrent le groupe socioprofessionnel (GSP) des individus retenus et son changement entre deux recensements successifs. Ils s'intéressent aux types de transition (stable, devenir actif, cesser l'activité, ...) ainsi qu'aux nouveaux venus (immigrés et naissances) et disparus (émigrés et décédés). L'analyse de statistique implicative donne une vision synthétique des liens entre ces dynamiques et les GSP concernés, ainsi qu'avec un certain nombre de variables démographiques et culturelles (sexe, âge, état-civil, religion).
L'objectif étant de mieux comprendre comment ces changements ou transitions sont liées aux caractéristiques démographiques des individus ainsi que leur impact sur la démographie de la Cité de Genève.

Les données étudiées ont été tirées des archives genevoises et proviennent plus particulièrement de six recensements de la population genevoise de 1816 1822, 1828, 1831,

1837 et 1843. Plus précisément, ils ont extrait que les informations relatives aux individus dont le patronyme commence avec la lettre "B", ce qui représentent environ 12,5% de la population. Au total sur les 6 recensements, ce ne sont pas loin de 30'000 notices individuelles qui ont été relevées.

L'analyse est centrée sur les recensements séparés de 6 ans. Ils ont considéré ainsi les états au temps t, soit 1816,1822, 1831 et 1837, et au temps t + 6 (1822, 1828, 1837, 1843), ainsi que l'évolution de ces états entre t et t + 6. Parmi les informations collectées, ils ont dénombré environ 1200 métiers qu'ils ont réorganisés d'une part en groupes socioprofessionnels, et d'autres parts en statuts sociaux. Le tableau 1 liste les catégories retenues pour chacun de ces regroupements. Pour distinguer les états au début de l'intervalle de ceux à la fin ils ont préfixé les notations par 't_' pour indiquer le début. Par exemple 't_gsp_hor' correspond à horloger en t et 'gsp_hor' à horloger en t + 6. Le tableau 2 indique comment les cas retenus se répartissent selon ces catégories au temps t.
Les transitions auquel ils s'intéressent sont définies à partir des groupes socioprofessionnels et sont récapitulées au tableau 3.

Statuts sociaux		Groupes socioprofessionnels	
ss_inc	Inconnu	*gsp_inac*	Inactif
ss_nqua	Manuel sans qualification	*gsp_nqua*	Sans qualification
ss_art	Manuel qualifié	*gsp_art*	Artisan
ss_colb	Col blanc	*gsp_hor*	Horloger
ss_pmb	Petite et moyenne bourgeoisie	*gsp_com*	Commerçant
ss_eli	Elites	*gsp_serv*	Services privés et publiques

TAB. 1 – *Liste des groupes socioprofessionnels et statuts sociaux*

Ils ont retenus également comme variables pour leur analyse l'âge, le sexe et l'état civil. Par ailleurs, intéressé par l'impact possible de la montée du catholicisme qui passe de 11% en 1816 à 23% en 1943, ils ont également inclus la religion dont ils distinguent trois modalités : protestant, catholique et autre. S'agissant de l'état civil, le nombre de divorcés étant très faible (< 10) ils ont considéré que les états célibataire, marié et veuf. Pour la religion comme pour l'état-civil ils ont retenu l'état au début t de l'intervalle censitaire à l'exception des nouveaux venus pour lesquels ils ne disposent que de l'état en t+6. Pour l'âge, ils ont retenu celui du milieu de l'intervalle, soit l'âge en t+3. Pour les analyses de statistique implicative, l'âge a été discrétisé en 3 classes (en minimisant la variance intra groupe) : age1 à moins de 16 ans, age2 de 16 à 41 ans et age3 pour les plus de 41 ans. Ils ont distingues

également deux périodes, l'une couvrant les transitions de 1816 à 1822 et de 1822 à 1828 et la deuxième celles de 1831 à 1837 et de 1837 à 1843.

Statuts / Groupes socioprofessionnels	Inconnu	Manuel ss qual.	Manuel qualifié	Col blanc	P.M.B.	Elite	Total
Inactif	4467	23	0	79	1	344	4914
Sans qualification	274	1672	96	118	3	0	2163
Horlogerie	0	71	1330	0	213	0	1614
Artisan, manuel qualifié	0	173	1527	3	80	0	1783
Commerce	0	112	64	21	537	7	741
Services publics et privés	0	28	18	37	156	82	321
Total	4741	2079	3035	258	990	433	11536

TAB. 2 – *Croisement des groupes socioprofessionnels avec les statuts sociaux (au temps t)*

Transition	(désignation)	GSP en t	GSP en $t+6$	autre condition
reste inactif	(inactif)	inactif	inactif	
devient actif	(nv_actif)	inactif	actif	
stable	(stable)	actif	actif	$\text{GSP}(t) = \text{GSP}(t+6)$
mobile	(mobile)	actif	actif	$\text{GSP}(t) \neq \text{GSP}(t+6)$
cesse l'activité	(retraite)	actif	inactif	
nouveau venu	(nv_venu)	non présent	présent	
disparu	(disparu)	présent	non présent	

TAB. 3 – *Les types de transitions*

Les données ont été examiné avec les outils de la statistique implicative, soit plus particulièrement ceux mis à disposition par le logiciel CHIC, à savoir l'arbre de similarité, l'arbre cohésitif et le graphe implicatif.

La figure 1.6 montre un premier graphe d'implication obtenu en n'incluant dans l'analyse que les transitions et les groupes socioprofessionnels d'origine et de destination. Le graphe a été obtenu avec des seuils relativement bas. Toutefois, cela ne remet pas en cause la pertinence statistique du graphe. En effet, comme ils ont utilisé la mesure entropique d'implication les seuils ne doivent pas être interprétés comme des significations statistiques. Avec la mesure classique non entropique, les relations indiquées ici sont d'ailleurs toutes significatives à des seuils supérieurs à 95%, mais se trouvent être noyées dans une quantité d'autres relations d'interprétation moins intéressante.

La figure 1.6, démontre l'efficience de la méthode, des relations triviales découlant des définitions du tableau 3, à savoir que ceux qui restent inactifs sont inactifs en t et en t +6, que les retraités deviennent inactifs en t+6 et que les nouveaux actifs étaient inactifs en t.

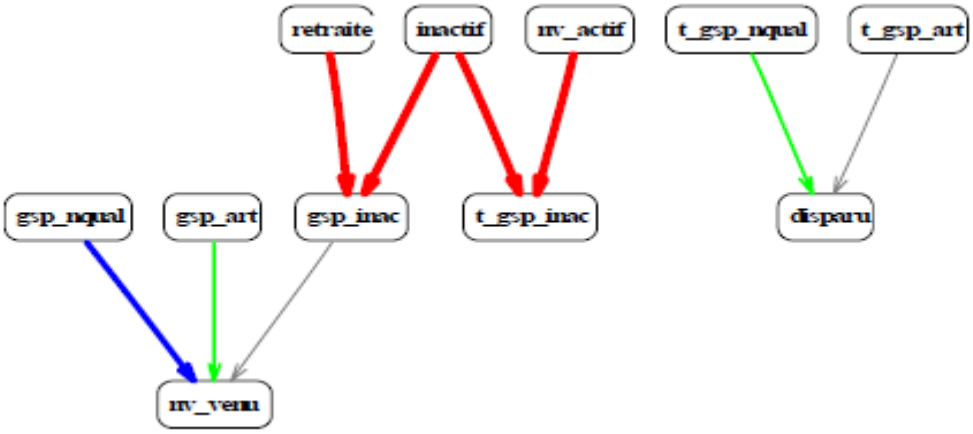

Figure 1.6: Transitions et groupes socioprofessionnels. Mesure entropique, seuils 99%, 81%, 63%, 58%.

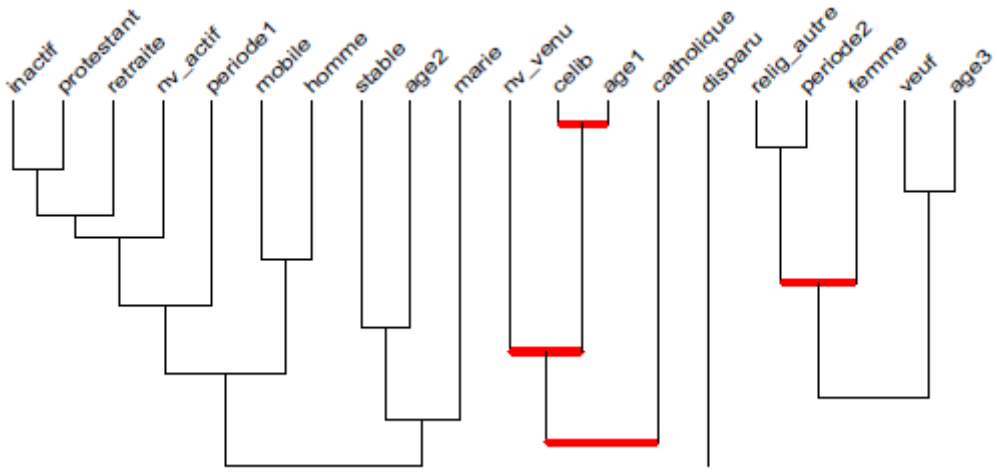

Figure 1.7: transition et caractéristiques dermographiques : arbres de similarités

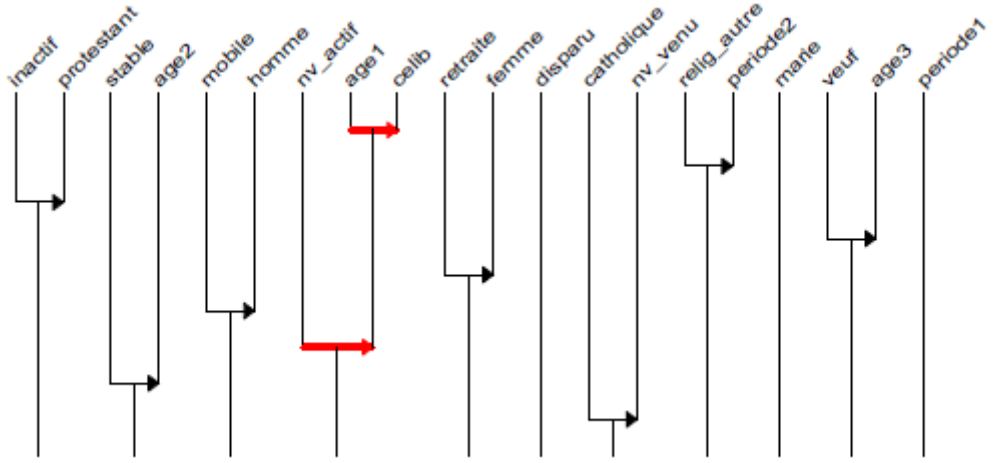

Figure 1.8: Transition et caractéristiques démographiques : arbre cohésitif, mesure entropique.

Les figures 1.6, 1.7 et 1.8 montrent respectivement l'arbre des similarités, l'arbre cohésitif et le graphe implicatif obtenus en incluant les transitions, les variables démographiques et la religion. L'arbre des similarités fait ressortir clairement trois groupes : en partant de la gauche, le premier correspond aux enracinés qui restent dans la cité, il comprend aussi bien les gens stables qui restent dans leur groupe socioprofessionnel que les mobiles, ou encore ceux qui deviennent actifs ou cessent leur activité. On voit que ces types de transitions se recoupent essentiellement avec les caractéristiques protestant, homme et marié. Le second comprend les nouveaux arrivants, et le troisième ceux qui quittent la cité. Les nouveaux arrivants semblent être plutôt jeunes et donc célibataires et catholiques. Les disparus ne se regroupent avec aucune caractéristique démographique.

Le graphe implicatif obtenu dans la figure 1.6 avec le critère entropique et des seuils relativement bas est plus intéressant. On y observe quatre modalités qui polarisent les effets des types de transition. Ce sont 'protestant', 'femme', 'homme' et 'célibataire'. Les nouveaux venus, les inactifs et ceux qui commencent une vie active sont principalement célibataires, les deux derniers groupes étant de jeunes célibataires. L'inactivité, le début d'activité et la cessation d'activité sont plus le fait des femmes, tandis que la stabilité et la mobilité socioprofessionnelle sont l'apanage des hommes. Quant à l'attribut 'protestant' il est associé à toutes les transitions internes, c'est-à-dire toutes sauf les nouveaux venus et les disparus. Ceci indique que si l'on trouve une majorité de protestants dans tous les groupes de transitions non migratoires, les protestants sont moins concernés par les mouvements migratoires.

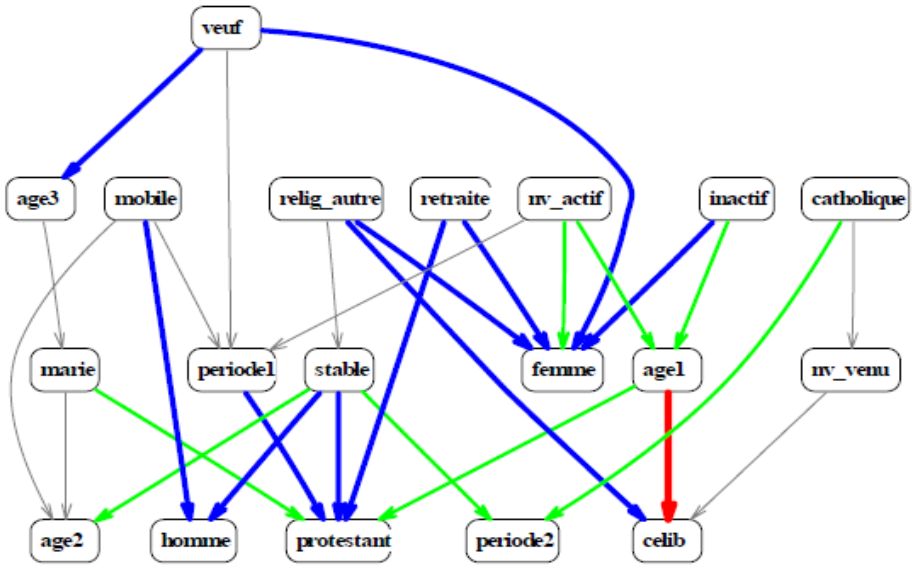

Figure 1.9 : Transitions et variables démographiques : graphe implicatif, mesure entropique, seuils 99%,75%, 65%,55%

Chemins / Statuts	Inconnu	Manuel ss qual.	Artisan, manuel qualifié	Col blanc	P.M.B.	Élite
stable ⇒ protestant	.	x	x	x	x	.
stable ⇒ homme	.	.	x	x	.	x
mobile ⇒ homme	.	.	x	x	x	x
nv_actif ⇒ protestant	.	x	x	x	x	.
nv_actif ⇒ célibataire	.	x	x	x	x	.
nv_actif ⇒ age1 ⇒ célibataire	x	.	.	x	.	.
nv_actif ⇒ femme	.	x	.	.	.	.

TAB. 4 – *Typicalité des statuts sociaux pour quelques chemins*

Par rapport à la figure 1.9, il est intéressant de noter que l'implication stable => protestant n'est pas avec des typicalités supérieures à 0.75. L'implication stable=> homme est, quant à elle, typique des artisans, des cols blancs et de l'élite, mais ne l'est ni de la petite et moyenne bourgeoisie, ni des non qualifiés. Il en va de même du fait que la mobilité socioprofessionnelle soit l'apanage des hommes, si ce n'est que cette dernière relation est également typique parmi la petite et moyenne bourgeoisie. Si l'on considère maintenant les implications importantes de 'nouvel actif' sur 'célibataire' et 'protestant', les indices de typicalités montrent qu'ils caractérisent les non qualifiés, les artisans, les cols blancs et la

petite et moyenne bourgeoisie, les statuts inconnus et les cols blancs étant typiques du chemin qui passe par la catégorie 'jeunes' (age1).

1.8.2. L'application de L'ASI pour l'analyse des résultats des étudiants : [B6]

Cette étude a été effectuée dans le but d'appliquer la méthode d'Analyse Statistique Implicative (ASI) aux notes des étudiants d'informatique de l'université A/Mira de Bejaia. L'ASI est utilisé pour découvrir et analyser les implications les plus pertinentes entre les différents modules de formation étudiés, nous allons présenter les résultats présentés dans le papier Khaled et al (2014) publié à ISKO-Maghreb (2014).

La population considérée représente les étudiants de licence 2 de département Informatique de l'université A/Mira de Bejaïa pendant les années 2010, 2011 et 2012. Les étudiants de licence 2 sont évalués sur les modules programmation linéaire (PL), théorie de langages (THL) logique mathématique (LogMat), analyse numérique (ANum), système d'exploitation (SE), probabilités et statistiques (P.S), traitement de signal (TSig), génie logiciel (GL), algorithme et structure de données (ALSTRD2), architecture (ARCH), bases de données (BDD) et structure de données (STRD1). Les notes des étudiants sont représentées par un fichier (un exemple est montré dans le tableau 4) qui contient 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 pour dire que les notes de modules sont à partitionner en un nombre fixe d'intervalles, grâce a l'utilisation de l'algorithme des nuées dynamiques Diday (1971) qui constitue automatiquement les intervalles qui ont des limites distinctes.

	ARCH p	STRD1 p
0809TMI02	11,25	11,5
09MI0034	9,38	11,38
09MI0590	10,63	9,38
…	…	…

Tableau 5 : Exemple de données

Les intervalles des notes des étudiants ici sont également partagés 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

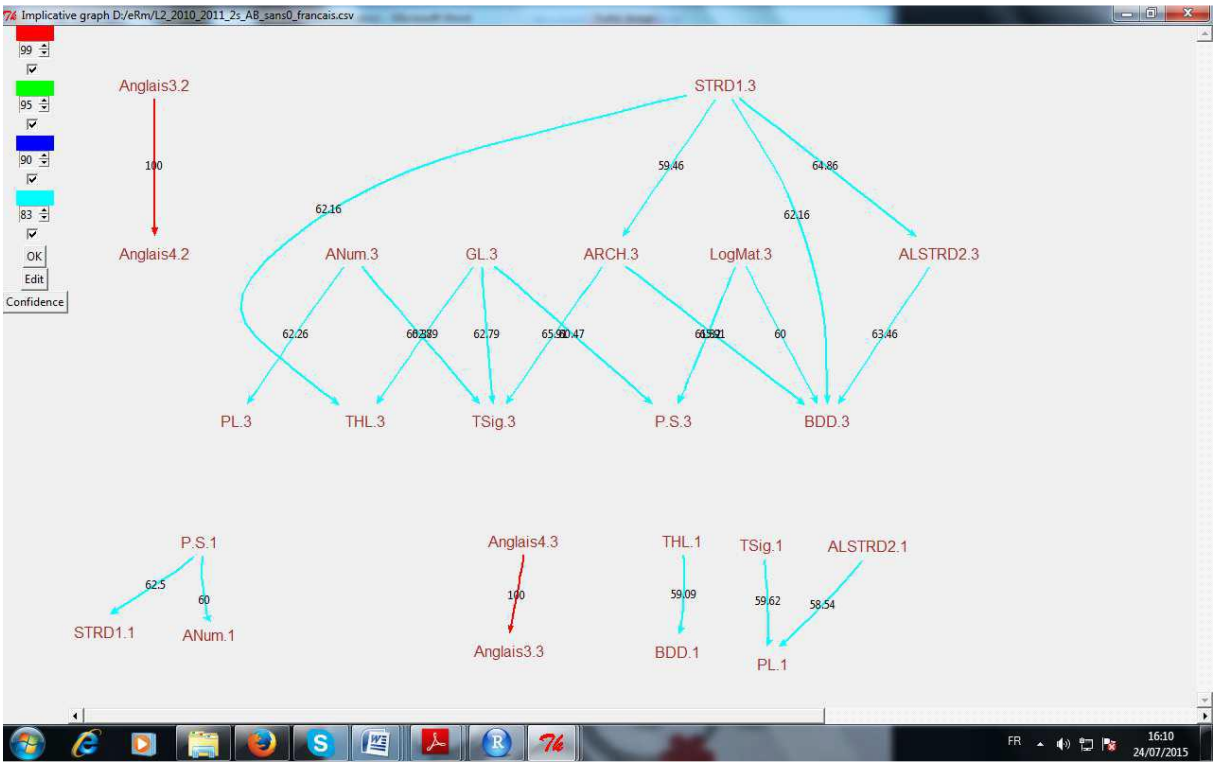

Figure 1.10 : Graphe implicatif L2 2010-2011

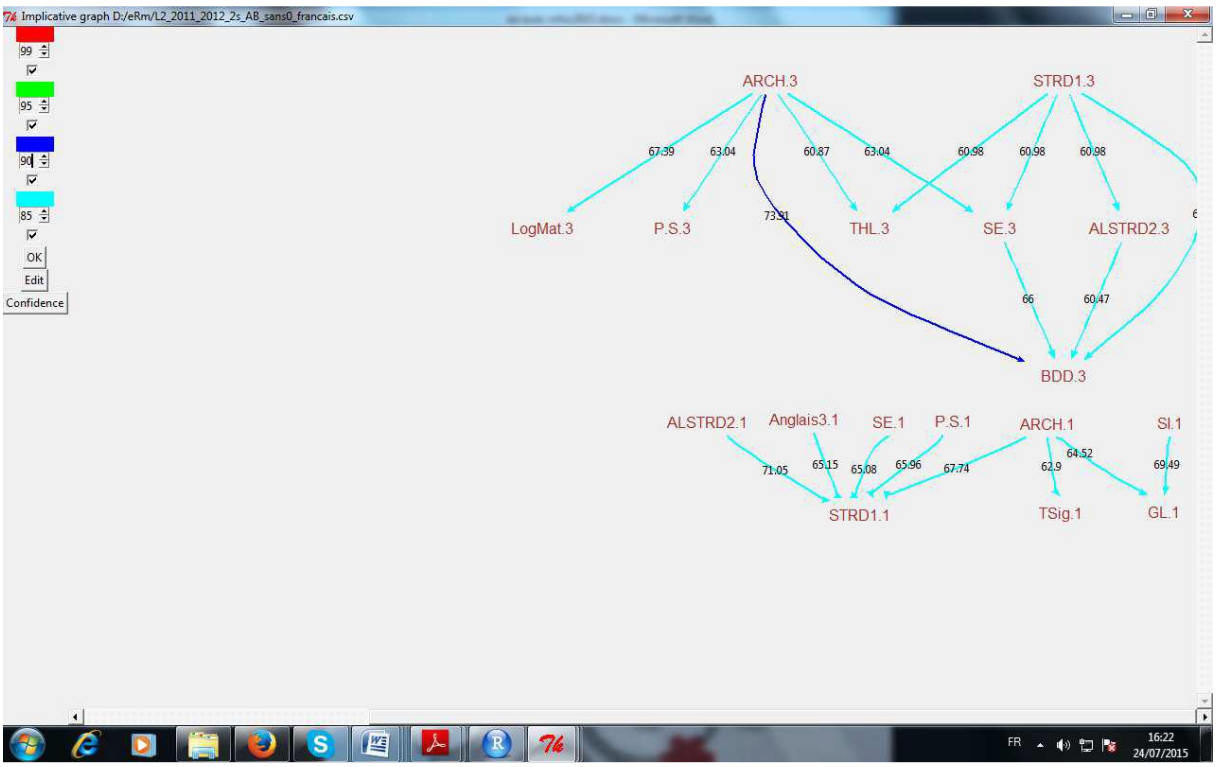

Figure 1.11 : Graph implicatif L2 2011-2012

Le critère utilisé pour la recherche des implications pour les graphes implicatifs est L'implifiance présenté dans GRAS *et al* (2015), ce dernier présente une combinaison entre les deux critères confiance et intensité d'implication. Nous voyons sur ces figures que les arcs sont pondérés par un poids ce qui représente la confiance, par exemple la confiance de 71,05 entre les modules ALSTRD2.1 et STRD1.1 dans la figure 1.10 ceci signifie que si un étudiant est faible en ALSTRD2 il a une chance de 71,05% d'être aussi faible en STRD1. Toutes les implications obtenues sont intéressantes car elles ont presque toutes des confiances supérieures à 60. Sur les deux figures 1.10 et 1.11 des implications qui se répètent sur deux années consécutives. Les auteurs ont remarqué clairement qu'un module réussi par les bons élèves entraine d'autres modules réussis par les bons élèves (par exemple STRD1.3 -> ALSTRD2.3, STRD1.3 -> THL3, ALSTRD2.3 -> BDD3). Il en est de même pour les modules échoués par les élèves faibles qui impliquent d'autres modules échoués par des élèves faibles (par exemple PS1 -> STRD1.1).sur les figures nous ne voyons pas des étudiants moyen cela est due au fait qu'un étudiant qui est moyen dans un module peut être faible ou bon dans d'autres modules et il n'est pas forcément moyen dans tous les modules.

1.8.3. L'ASI au service d'une étude sur l'anticipation des départs à la retraite : [B7]

Dès la réforme des retraites en juillet 2008 en France un certain nombre d'entreprises ont constaté que celle-ci entraînait des modifications importantes du comportement des agents ou des employés. Cette reforme permet aux agents de prolonger ou non leur activité, Une mesure phare de la réforme est en effet la suppression de la mise à la retraite d'office. Elle implique que les agents peuvent dorénavant choisir leur date de départ. La gestion des ressources humaines a donc commandé une étude statistique afin de déterminer les variables en jeu dans ces décisions. Cette étude menée par Jean-Claude Oriol & Anicée Chancel concerne la SNCF (Société nationale des chemins de fer français) exactement elle prend en compte que les agents pouvant partir à la retraite entre le 1er janvier 2008 et le 31 décembre 2009.

Les variables utilisées sont les suivantes : Situation personnelle (genre, situation familiale Interruption de Longue Durée (ILD), région), Qualification, Ancienneté à la SNCF Pourcentage d'utilisation, Activité (Catégorie (roulant ou sédentaire), Branche, Pénibilité).

Moins_de_30ans : ancienneté inférieure à 30 ans à la SNCF

Branche_PS : Branche protection sociale

Moins_de_70% : Temps partiel inférieur à 70%

Branche_FT : Branche fonctions transverses

PACA : Région Provence-Alpes-Côte d'azur

Branche_DCF : Branche direction de la circulation ferroviaire

70_a_100% : Temps partiel supérieur ou égal à 70%

Natild_Autre : Nature d'interruption de longue durée autre que détaché ou maladie.

L'arbre de similarités obtenues est le suivant :

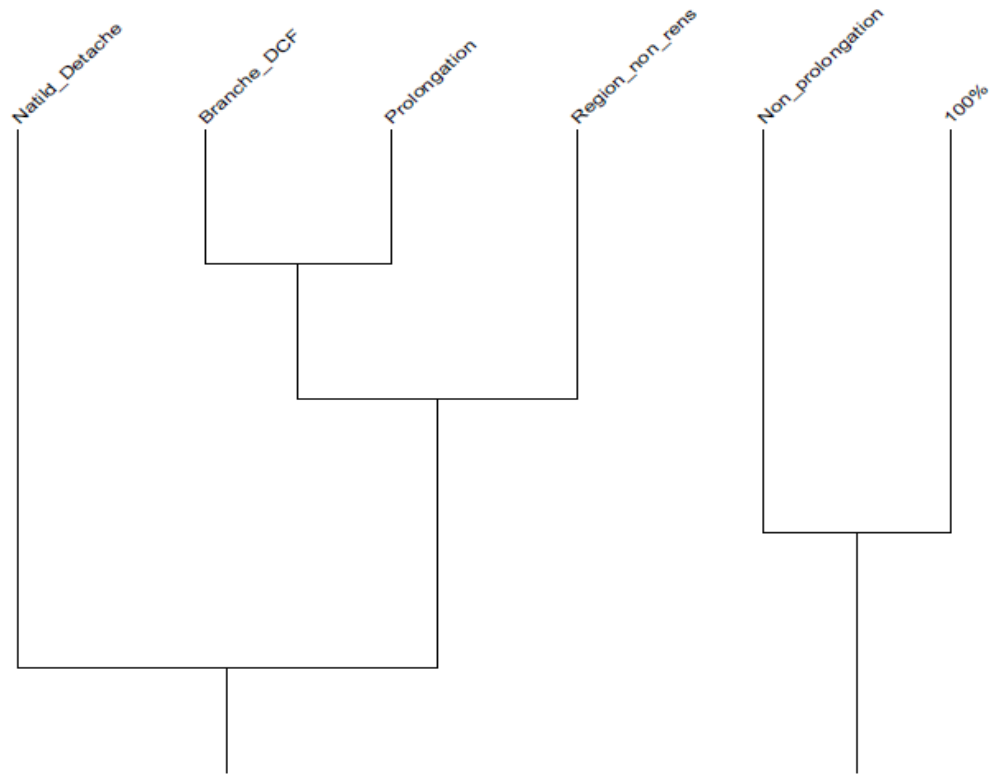

Figure 1.12: Arbre des similarités

Les variables « non prolongation » et « 100% » sont similaires avec un Indice de similarité égale a 0,955217 C'est-à-dire qu'il y a beaucoup d'agents à temps plein et qui ne prolongent pas leur activité De même, les variables « branche Direction des Circulations Ferroviaires (branche_DCF) » et « prolongation » sont similaires avec un IDS(Indice de similarité) de 1 la variable « région non renseignée (region_non_rens)» leur ressemble avec un IDS de 0,99997. Au niveau suivant (IDS: 0,925657), la variable « détaché » est similaire à ces trois variables.

L'arbre cohésitif obtenu est le suivant :

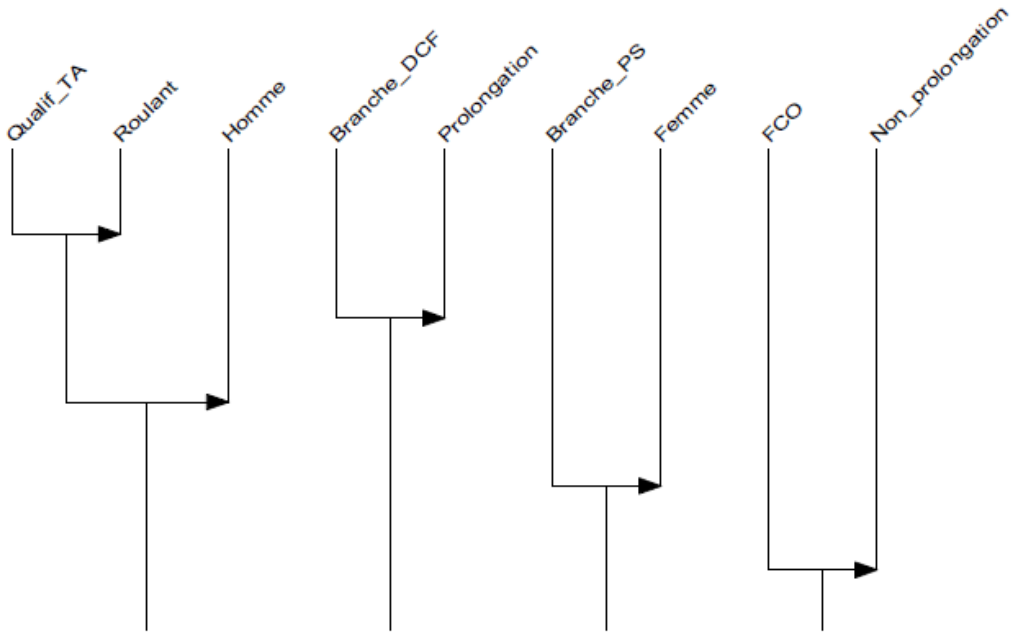

Figure 1.13 : Arbre cohésitif

L'emploi dans la branche Direction des Circulations Ferroviaires implique la prolongation d'activité avec un critère de cohésion de 1. De plus le fait de travailler dans la région Franche-Comté (FCO) implique la non prolongation d'activité mais avec un critère de cohésion de seulement 0,532. L'emploi dans la branche Protection Sociale (branche_PS) implique le fait d'être une femme (critère de cohésion : 0,99). Enfin, un emploi de qualification TA implique un emploi de roulant (critère de cohésion : 1). Au niveau suivant (critère de cohésion : 1) ces variables impliquent le fait d'être un homme.

Les graphes implicatifs obtenus sont les suivants :

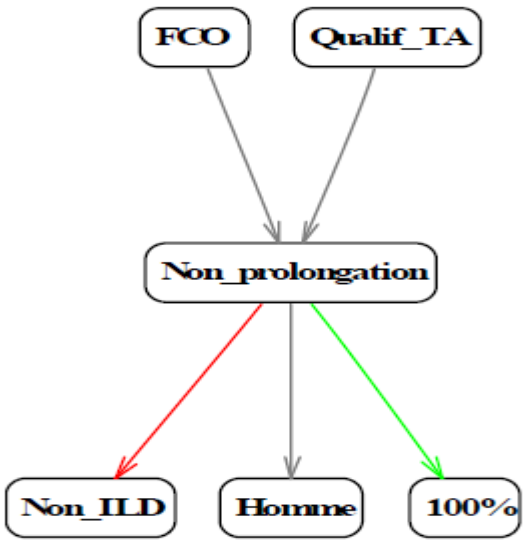

Figure 1.14 : Graphe implicatif obtenu aux seuils 0.95 (en rouge), 0.90 (en bleu), 0.80 (en vert) et 0.70 (en gris) en ne sélectionnant que les chemins dans lesquels la variable non prolongation est présente

Les agents travaillant dans la région Franche-Comté et les agents de qualification TA ne prolongent pas leur activité. De plus, les agents qui ne prolongent pas leur activité sont des hommes, sont à temps plein et ne sont pas en situation d'interruption de longue durée(Non_ILD).

Figure 1.15 : Graphe implicatif obtenu aux seuils 0.95 (en rouge), 0.90 (en bleu), 0.80 (en vert) et 0.70 (en gris) en ne sélectionnant que les chemins dans lesquels la variable prolongation est présente.

D'après le dernier graphe implicatif les agents des branches protection sociale, fonctions transverses et direction de la circulation ferroviaire prolongent leur activité. Les agents en situation d'interruption de longue durée, quelle qu'en soit la raison (détaché, maladie ou autre) prolongent eux aussi leur activité la même chose concernant les personnes travaillant dans les régions Ile de France, Provence-Alpes-Côte d'Azur, Alsace et Lorraine et les agents de qualification F à H(les cadres). Les femmes, les agents dont l'ancienneté à la SNCF est inférieure à 30 ans et les agents employés à temps partiel prolongent eux aussi leur activité, la branche protection sociale est principalement composée de femmes. Dans cette étude l'ASI a montré les paramètres qui influencent a la prise de décision de la prolongation d'activité.

1.9. Conclusion :

Dans ce chapitre nous avons présenté la méthode d'analyse de données non symétrique (ASI) son concept, nous avons présenté d'autres indices d'implication et les comparé avec l'intensité d'implication, ainsi quelques domaines d'application dans le chapitre suivant nous allons présenter le logiciel CHIC.

.

Chapitre 2

L'environnement de travail

2.1. Introduction :

CHIC est un logiciel d'analyse de données utilisé pour classifier, structurer, hiérarchiser, découvrir des régularités, des invariants, des similarités, des implications entre comportements. Dans ce chapitre nous allons présenter le logiciel CHIC, les données traitées, ces fonctionnalités et les étapes d'installation.

2.2. CHIC :

Un logiciel dénommé C.H.I.C. (acronyme de Classification Hiérarchique Implicative et Cohésitive) C'est un logiciel d'analyse de données, initialement conçu par Régis Gras en ce qui concerne les algorithmes, puis successivement développé sur P.C. par Saddo Ag Almouloud, Harrisson Ratsimba-Rajohn et, dans sa version actuelle, par Raphaël Couturier Professeur à l'Université de Franche-Comté, le logiciel est fonctionnel pour traiter tous les problèmes numériques et graphiques nécessaires à l'usage de la méthode A.S.I. Evolutif suivant les extensions de la théorie, Grâce à cet outil informatique, de nombreuses applications dans des domaines variés (pédagogie, psychologie, sociologie, bio-informatique, médecine, histoire de l'art, etc.) ont montré la pertinence du modèle probabiliste choisi et, de ce fait, la richesse des informations obtenues par le traitement des variables. Elles ont, en particulier, permis de mettre en évidence des pépites de connaissance que d'autres méthodes n'extrayaient pas et d'exprimer des relations causales affectées d'une intensité. Ses avantages tiennent à sa sensibilité aux effectifs des instances et aux représentations aisées à interpréter.

Des représentations graphiques permettent de visualiser les relations extraites du corpus donné et en structurent leur ensemble à travers la création de graphes implicatifs, d'arbres de similarités et d'arbres de cohésitif.[B8]

Ce Logiciel a pour fonctions essentielles d'extraire automatiquement des données, un ensemble de règles d'association quasi implicatives entre les variables (on Récupère la table complète des indices d'implication entre toutes les variables dans un fichier.csv), De fournir un indice de qualité de l'association, de représenter une structuration des variables obtenue au moyen de ces règles et de mettre en évidence la contribution des sujets ou des descripteurs de ces sujets par rapport aux règles ou aux familles de règles.[B8]

CHIC est le fruit informatique des travaux sur l'analyse statistique implicative. Une version de ce logiciel a été portée en C++ sous Windows il y a 10 ans environ a partir d'une version antérieure en Pascal, mais avec des développements importants et avec une plus grande convivialité Couturier (2000). Depuis elle a subi régulièrement de nombreuses modifications tant au niveau pratique que sur le plan théorique en intégrant de nombreux nouveaux modes de calculs et de nouveaux concepts. [B8], la version la plus récente est en R qui est appelé RCHIC c'est la version qui nous intéresse et avec laquelle nous allons travailler.

2.3. Les données traitées :

Initialement CHIC a été conçu pour gérer les variables binaires. Plus tard CHIC a été renforcé par d'autres types de variables. Actuellement, CHIC permet à l'utilisateur de gérer des variables binaires, modale et fréquentielle, quantitative ou intervalle. De plus elles peuvent être principales, c'est-à-dire qu'elles interviennent directement dans tous les calculs ou elles peuvent être secondaires comme il est fait en analyse factorielle. Les variables modales et fréquentielles doivent avoir une valeur réelle comprise entre 0 et 1. L'utilisateur doit faire attention à la façon dont les variables réelles sont transformées en variables de fréquence. Plusieurs stratégies sont disponibles en fonction des valeurs. Si les valeurs sont positives, elles peuvent être divisées par la valeur maximale. Une autre possibilité réside dans la considération que la valeur minimale représente 0 et le maximum représente 1, tous les autres variables sont réparties proportionnellement entre le minimum et les valeurs maximales. Si une variable réelle a des valeurs positives et négatives, il est possible de diviser les variables en deux variables, une pour les valeurs positives et un autre pour les

valeurs négatives. Cependant, il est possible de considérer que le minimum (même si elle est négative) représente 0 et le maximum représente 1. Dans ce cas, toutes les autres valeurs sont transformées dans l'intervalle [0, 1].Les valeurs des variables quantitatives sont normalisées dans l'intervalle [0-1] en divisant toutes les valeurs par la valeur maximum obtenue par la variable. Les variables-intervalles sont automatiquement découpées en différents intervalles par un algorithme approprié de type « nuées dynamiques » qui, à partir d'un nombre d'intervalles choisi par l'utilisateur, constitue des intervalles tout en maximisant la variance inter-classe. Ce type de variable est utilisé pour modéliser des situations plus complexes. [B8]

2.4. Les fonctionnalités :

Les fonctionnalités de CHIC sont multiples nous citons quelques-unes :

- Tout d'abord, il fournit les statistiques brutes : moyennes, écart-types, coefficients de corrélation
- Une classification hiérarchique des similarités selon l'algorithme de la vraisemblance du lien de I.C.Lerman : valeurs des similarités, arbre hiérarchique avec son processus et ses niveaux significatifs.

- une analyse implicative selon les méthodes de R. Gras , avec en option la méthode classique et la méthode entropique, tout en donnant la possibilité de conjoindre, disjoindre, supprimer, ajouter des variables
- Les valeurs d'intensité d'implication ou de similarité, les coefficients de corrélation linéaire, les croisements deux à deux des variables,
- Le graphe implicatif selon différents seuils d'intensité, les sujets et les catégories de sujets contribuant aux chemins du graphe, les chemins significatifs (travaux de M.Bailleul)
- La classification cohésitive en un arbre, avec ses niveaux significatifs, la contribution de sujets et catégories de sujets à ces niveaux
- La réduction du nombre de variables grâce à un sous-programme qui permet de traiter des gros fichiers tout en gardant le maximum d'information relativement à la similarité ou à l'implication. [W1]

A l'initialisation de chaque calcul, CHIC indique à l'utilisateur des informations telles que le nombre d'occurrences, la moyenne et l'écart-type de chaque variable. Lorsqu'on analyse des données contenant des grands effectifs, l'option calcul entropique est préférée à l'option calcul classique car le calcul d'une règle prend également en compte la contraposée de celle-ci. Ainsi la qualité de la règle s'en trouve renforcée. [W1]

2.5. Installation de RCHIC :

2.5.1. Installation de logiciel R :

Pour une installation sous Windows**,** on se rendra sur cette page : http://cran.r-project.org/bin/windows/base/ et l'on suivra le premier lien pour télécharger le programme d'installation. Une fois le programme d'installation lancé, il suffira d'installer **R** avec les options par défaut.

2.5.1.1. Présentation de R :

Le logiciel R est un logiciel de statistique crée par Ross Ihaka & Robert Gentleman, R est open source Il permet, entre autres, de réaliser des analyses statistiques. Plus particulièrement, il comporte des moyens qui rendent possibles la manipulation des données, les calculs et les représentations graphiques. **R** a aussi la possibilité d'exécuter des programmes stockés dans des fichiers textes. [B9]

En effet **R** possède :

- un système efficace de manipulation et de stockage des données,
- différents opérateurs pour le calcul sur tableaux (et spécialement les matrices),
- un grand nombre d'outils pour l'analyse des données et les méthodes statistiques,
- des moyens graphiques pour visualiser les analyses. [B9]

L'utilisation de **R** présente plusieurs avantages :

• c'est un logiciel multiplateforme, qui fonctionne aussi bien sur des sytèmes Linux, Mac OS X ou Windows

• c'est un logiciel libre, développé par ses utilisateurs et modifiable par tout un chacun
• c'est un logiciel gratuit
• c'est un logiciel très puissant, dont les fonctionnalités de base peuvent être étendues à l'aide de plusieurs milliers d'extensions
• c'est un logiciel dont le développement est très actif et dont la communauté d'utilisateurs ne Cesse de s'élargir
• les possibilités de manipulation de données sous **R** sont en général largement supérieures à Celles des autres logiciels usuels d'analyse statistique. [B9]

Le logiciel R est un logiciel en lignes de commande. Pour se servir de R, il faut taper des commandes dans une fenêtre. Son interface graphique de base est plutôt rudimentaire (voir figure ci-après).

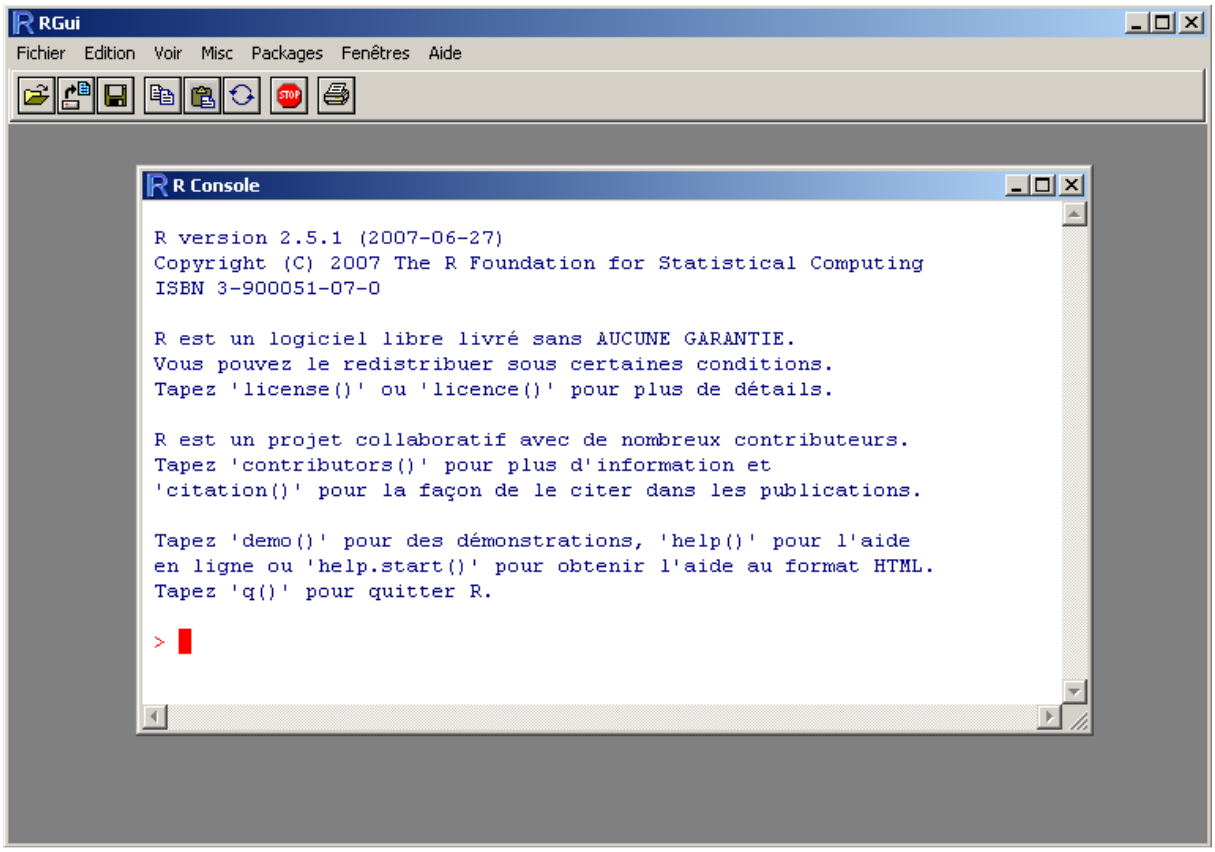

Figure 2.1: Interface de R sous Windows

R est un logiciel de type modulaire. On installe une version de base qui comprend un répertoire étendu de commandes relatives à la manipulation de fichiers de données, à des fonctions statistiques et graphiques et à des opérations de programmation. L'usager peut

ensuite enrichir cette version en puisant dans une banque de modules appelés *packages*. Ces modules couvrent une extraordinaire variété de techniques d'analyse dans des disciplines fort variées, Au moment où ces lignes sont écrites, il y a 7176 packages disponibles sur le site officiel. Le logiciel R dépasse de très loin les ressources des logiciels commerciaux d'analyse statistique. [B9]

2.5.1.2. Le langage R :

R et un langage de programmation (de script) interprété dérivé de S (disponible dans le logiciel S-PLUS). A ce titre, il en intègre toutes les caractéristiques : données simples et structurées, opération d'entrée-sortie, branchements conditionnels, boucles indicées et conditionnelles, récursivité, etc. R est largement utilisé par les statisticiens, les data miners, data scientists pour le développement de logiciels statistiques et l'analyse des données.[B9]

2.5.2. Installation de RStudio :

Une fois R correctement installé, rendez-vous sur :
http://www.rstudio.com/products/rstudio/download/ pour télécharger la dernière version stable de RStudio**.** Plus précisément, il s'agit de l'édition Open Source de RStudio Desktop (en effet, il existe aussi une version serveur).

2.5.2.1. Présentation de RStudio :

RStudio est un environnement de développement intégré libre, gratuit, et qui fonctionne sous Windows, Mac OS X et Linux. Il complète R et fournit un éditeur de script avec coloration syntaxique, des fonctionnalités pratiques d'édition et d'exécution du code (comme l'autocomplétion), un affichage simultané du code, de la console R, des fichiers, graphiques et pages d'aide, une gestion des extensions, une intégration avec des systèmes de contrôle de versions comme git, etc. Il intègre de base divers outils comme par exemple la production de rapports au format Rmarkdown. Il est en développement actif et de nouvelles fonctionnalités sont ajoutées régulièrement. [B9]

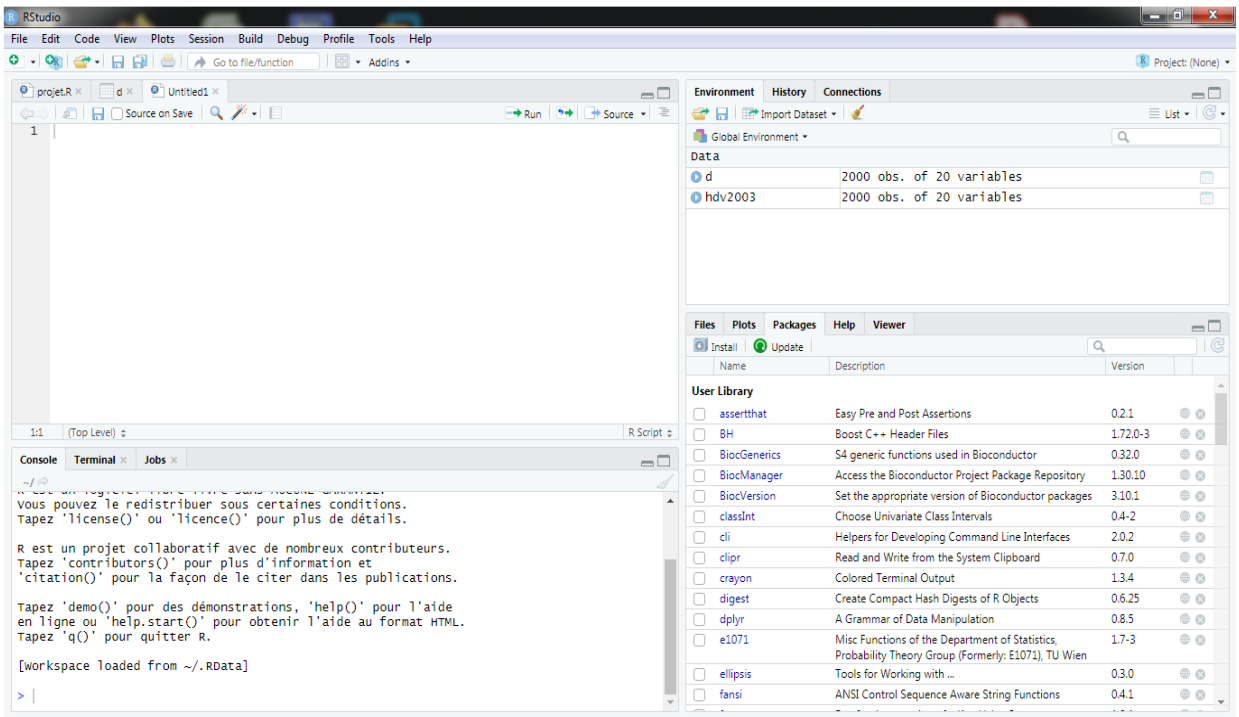

Figure 2.2 : Interface de RStudio sous Windows

Lorsque plusieurs versions de R sont disponibles, RStudio choisit par défaut la plus récente. Il est possible de spécifier à RStudio quelle version de R utiliser via le menu Tools > Global Options > General.

L'interface de RStudio est divisée en quatre quadrants : [B9]

- ✓ Le quadrant supérieur gauche est dédié aux différents fichiers de travail
- ✓ le quadrant inférieur gauche correspond à ce que l'on appelle la console, c'est-à-dire à R proprement dit.
- ✓ le quadrant supérieur droit permet de connaître :
  - La liste des objets en mémoire ou environnement de travail (onglet Environment)
  - Ainsi que l'historique des commandes saisies dans la console (onglet History)
- ✓ le quadrant inférieur droit affiche :
  - la liste des fichiers du répertoire de travail (onglet Files),
  - les graphiques réalisés (onglet Plots),
  - la liste des extensions disponibles (onglet Packages),
  - l'aide en ligne (onglet Help)

- un Viewer utilisé pour visualiser certains types de graphiques au format web.

2.5.3. Installation de Rchic : [W2]

Une fois R et RStudio sont installés on passe maintenant a l'installation de Rchic.

Rchic est un package pour R qui implémente la plupart des outils de l'analyse statistique implicative, qui implémente le graphe de similitude, le graphe hiérarchique et le graphe implicatif.

Les utilisateurs peuvent télécharger le projet Rchic dans github :

https://github.com/rcouturier/Rchic

Pour utiliser rchic, certains packages sont requis:

Rgraphviz : fournit des capacités de traçage pour les objets du graphique R

pour installer ce package entrez dans la console ce qui suit :

```
source ("https://bioconductor.org/biocLite.R")
biocLite ("Rgraphviz")
```

Stringr : ce package fournit les fonctions de manipulation de chaînes les plus importantes et les plus utilisées, pour l'installer entrez dans la console ce qui suit :

```
install.packages ("stringr")
```

tcltk2 : Extension de la bibliothèque tcltk(Bibliothèque disponible dans R qui contient les principales commandes Tcl et widgets Tk) qui permet d'accéder à des styles et des fonctions supplémentaires. Pour l'installer on utilise la même manière que le package précédent.

Rcpp : est un package d'extension pour R qui offre une interface fonctionnelle entre C ++ et R. Cependant, elle est quelque peu différente d'un package R traditionnel car son composant clé est une bibliothèque C ++.

L'importation et l'utilisation de fonctions codées en C++ dans R se fait grâce au package Rcpp. Pour l'installer on utilise la même manière que le package précédent.

Après l'installation, vous devez charger la bibliothèque rchic (library) puis vous pouvez utiliser rchic avec la commande rchic() :

```
Library (rchic)
rchic()
```

Apres l'exécution de la commande rchic() une fenêtre s'apparaître comme la montre la figure suivante :

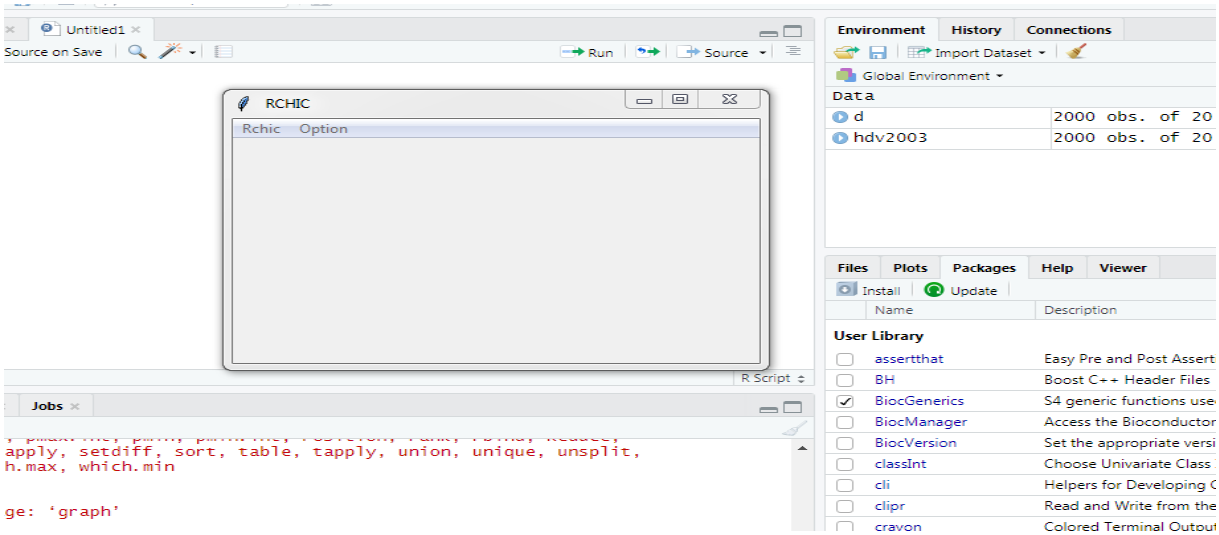

Figure 2.3 : fenêtre rchic

le menu rchic permet de choisir le calcul souhaité (graphe implicative,arbre similarité ou hiéarchique) Ensuite, vous devez utiliser un fichier .csv

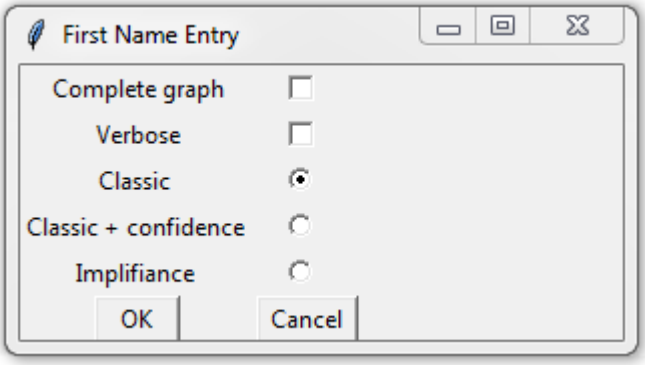

Figure 2.4: les options possible avec CHIC

Dans les options de CHIC, nous pouvons choisir d'utiliser l'implication classique ou l'implication entropique.

2.6. Conclusion :

Le logiciel CHIC, RCHIC permet d'effectuer différents traitements statistiques basé sur l'étonnement statistique (analyse des similarités ou analyse implicative). Nous avons essayé d'uniformiser les différentes techniques ou options. Dans ce chapitre nous avons détaillé les fonctionnalités que nous pouvons trouver sur ce logiciel ainsi les étapes de son installation.

Chapitre 3

Application de l'analyse statistique implicative

3.1. Introduction :

Dans ce chapitre nous allons appliquer L'ASI a des données qui concerne la maladie du cancer de sein en utilisant deux options (l'amplifiante et l'implication classique associé a la confiance), Nous allons analyser et comparer les résultats.

3.2. Présentation des données traitées :

Les données que nous avons utilisé concernent la maladie du cancer du sein (WBC, WDBC) qui ont été obtenues auprès des hôpitaux de l'Université du Wisconsin, Madison, États-Unis auprès du Dr William H. Wolberg, W. Nick Street(Département des sciences informatiques University of Wisconsin), et Olvi L. Mangasarian(Computer Sciences Dept University of Wisconsin). La population considérée représente les cas cliniques rapporté par Dr Wolberg, 699 instances traitées pour le cas de WBC, 570 pour WDBC. Les variables utilisés pour les deux jeux de données décrivent les caractéristiques des noyaux cellulaires ainsi deux variables de types binaires, Ces dernières nous permettent de connaitre si la personne est atteinte ou pas du cancer. Un exemple de données traitées (WBC) est dans la figure 3.1, Dans le jeu de données chaque variable est suivie de la lettre « p », cela pour

dire que ces variables sont a partitionner en un nombre fixe d'intervalles, le nombre de partition par défaut pour les variables est 3 par exemple pour la variable V1, V1.1 veut dire les valeurs faibles de la variable V1, V1.2 les valeurs moyennes et V1.3 ce sont les valeurs fortes de la variable V1. Donc nous allons nous intéresser à toutes les implications dont la conclusion est maline ou bénine (atteint du cancer ou pas). [W3]

V1 p	V2 p	V3 p	V4 p	V5 p	V6 p	V7 p	V8 p	V9 p	malignant	benign
5	1	1	1	2	1	3	1	1	0	1
5	4	4	5	7	10	3	2	1	0	1
3	1	1	1	2	2	3	1	1	0	1
6	8	8	1	3	4	3	7	1	0	1
4	1	1	3	2	1	3	1	1	0	1
8	10	10	8	7	10	9	7	1	1	0
1	1	1	1	2	10	3	1	1	0	1
2	1	2	1	2	1	3	1	1	0	1
2	1	1	1	2	1	1	1	5	0	1
4	2	1	1	2	1	2	1	1	0	1
1	1	1	1	1	1	3	1	1	0	1
2	1	1	1	2	1	2	1	1	0	1
5	3	3	3	2	3	4	4	1	1	0
1	1	1	1	2	3	3	1	1	0	1
8	7	5	10	7	9	5	5	4	1	0
7	4	6	4	6	1	4	3	1	1	0
4	1	1	1	2	1	2	1	1	0	1
4	1	1	1	2	1	3	1	1	0	1
10	7	7	6	4	10	4	1	2	1	0
6	1	1	1	2	1	3	1	1	0	1
7	3	2	10	5	10	5	4	4	1	0

Figure3.1 : Un Extrait dans le jeu de données du cancer du sein WBC

3.3. Application de l'ASI :

Cette partie est consacrée a examiner les jeux de données avec les outils de la statisque implicative, soit plus particulièrement ceux mis à disposition par le logiciel RCHIC, après diverses simulations utilisant les possibilités de l'ASI implémentées dans RCHIC. Nous avons choisi d'utiliser l'implication classique associée à la confiance et l'implifiance.

3.3.1. Utilisation de l'implication classique associée à la confiance :

Cette option nous permet d'obtenir les implications en utilisant l'indice classic qui est l'intensité d'implication avec la possibilité de rajouter un autre indice qui est la confiance, nous aurons les implications les plus étonnantes et les plus confiantes.

3.3.1.1. Application au jeu de données WBC :

La figure 3.2 montre le graphe implicatif obtenu avec un seuil d'implication de 95 et une confiance de 95%, à vrai dire les valeurs affichées représentent les valeurs de confiance.

Nous voyons que tout les malades qui possèdent des valeurs faibles des variables : V2, V7, V3 et V1 ne sont pas atteints du cancer du sein par contre ceux qui possèdent des valeurs fortes des variables : V6, V1, V3, V2, V8, V4 sont atteints du cancer du sein.

La figure 3.3 montre aussi un graphe d'implication obtenu avec un seuil d'implication de 80 et une confiance de 85%, nous voyons bien que les personnes qui ne sont pas atteintes du cancer portent les mêmes variables que le graphe précédent plus la valeur faible de la variable V6, et pour les personnes atteintes du cancer portent aussi les mêmes variables que le graphe précédent plus les valeurs fortes des variables : V7,V5 et V9 et la valeurs faible de la variable V9.

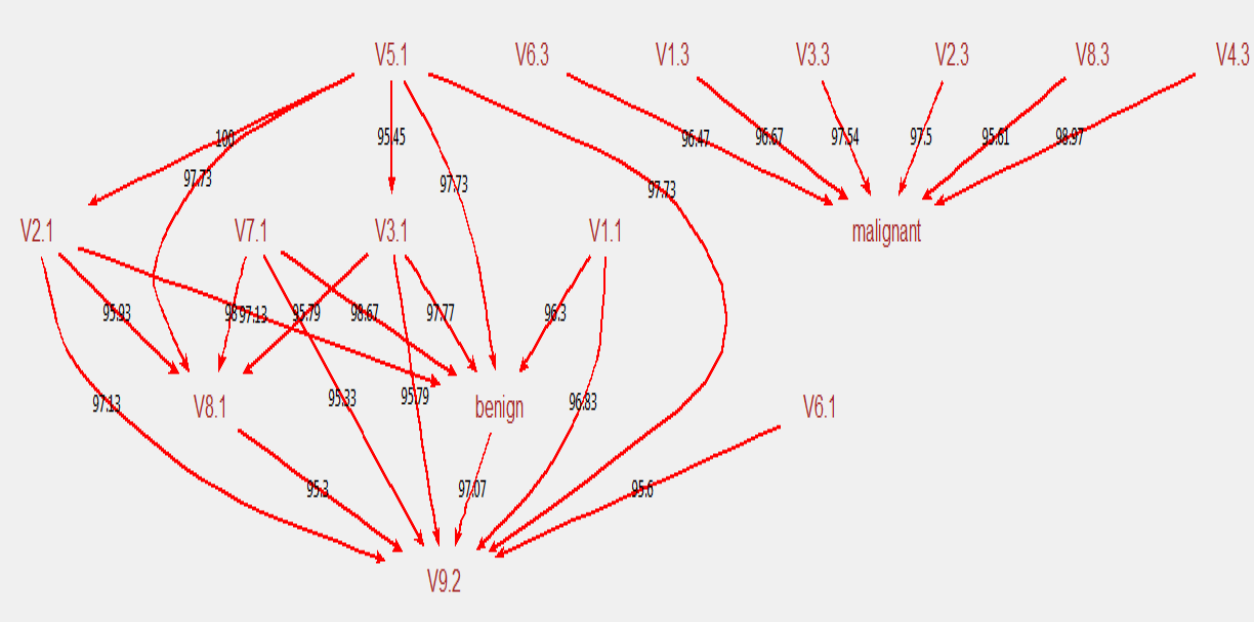

Figure 3.2: graphe implicatif, Seuils : intensité d'implication 95, confiance 95%

La figue 3.4 montre un graphe implicatif obtenu avec un seuil d'implication de 70 et une confiance de 70% pour les deux cas (atteints ou pas). les même résultats obtenus dans les deux graphes précédent sont obtenus dans ce graphe mais avec l'ajout de quelques variables, pour les personnes atteintes du cancer, portent des valeurs moyennes des variable : V3,V2,V8 et pour ceux qui sont pas atteints portent des valeurs faibles des variables : V5,V1.donc d'après ces trois graphes nous pouvons conclure que les valeurs faibles des variables conduisent a un état sein et les valeurs fortes conduisent a un cancer du sein. Les valeurs fortes et faibles de la variable V9 conduisent a un cancer du sein, ce qui fait que cette variable nous donne aucune indication sur l'état de l'individu (soit malin ou bénin).

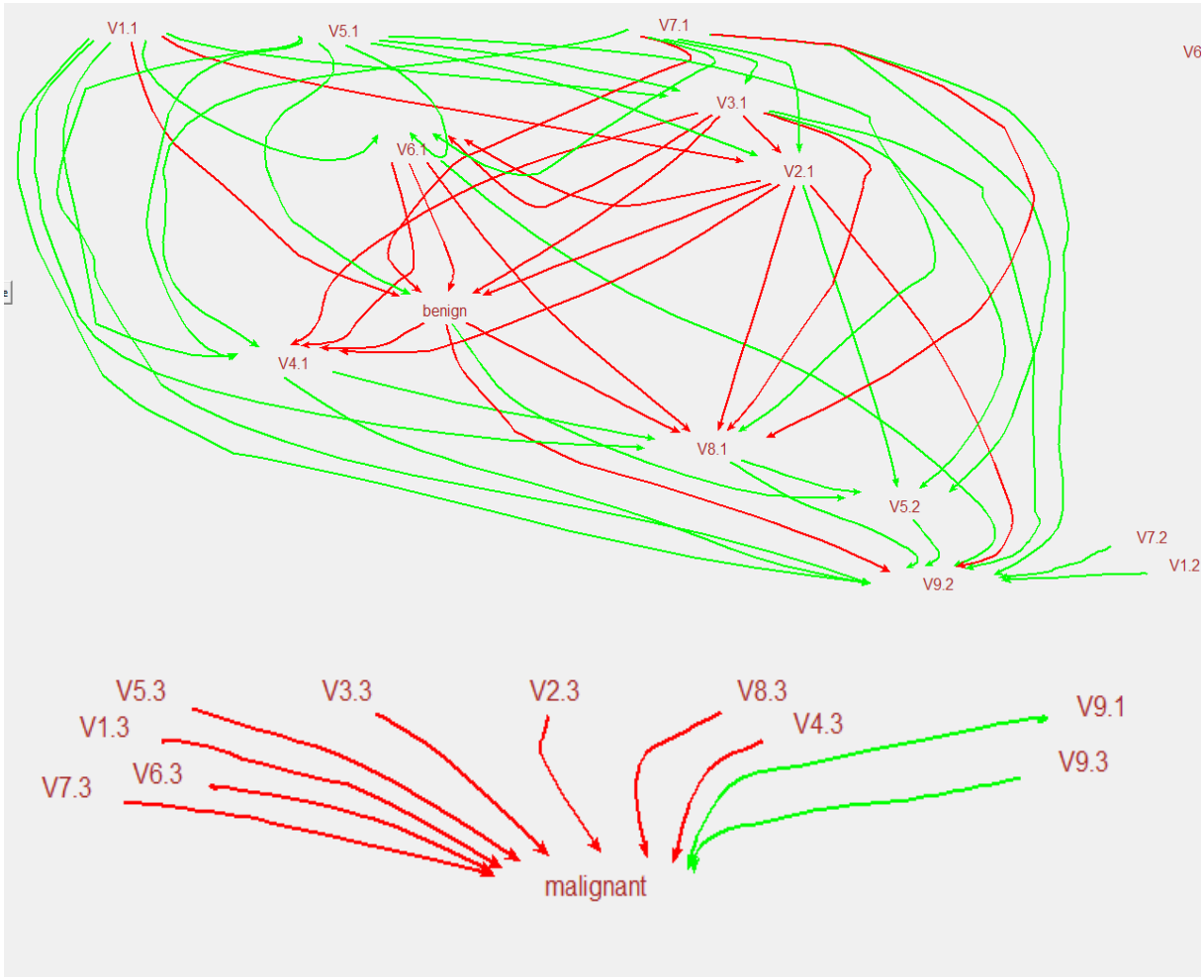

Figure 3.3 : graphe implicatif, Seuils : intensité d'implication 80, confiance 85%

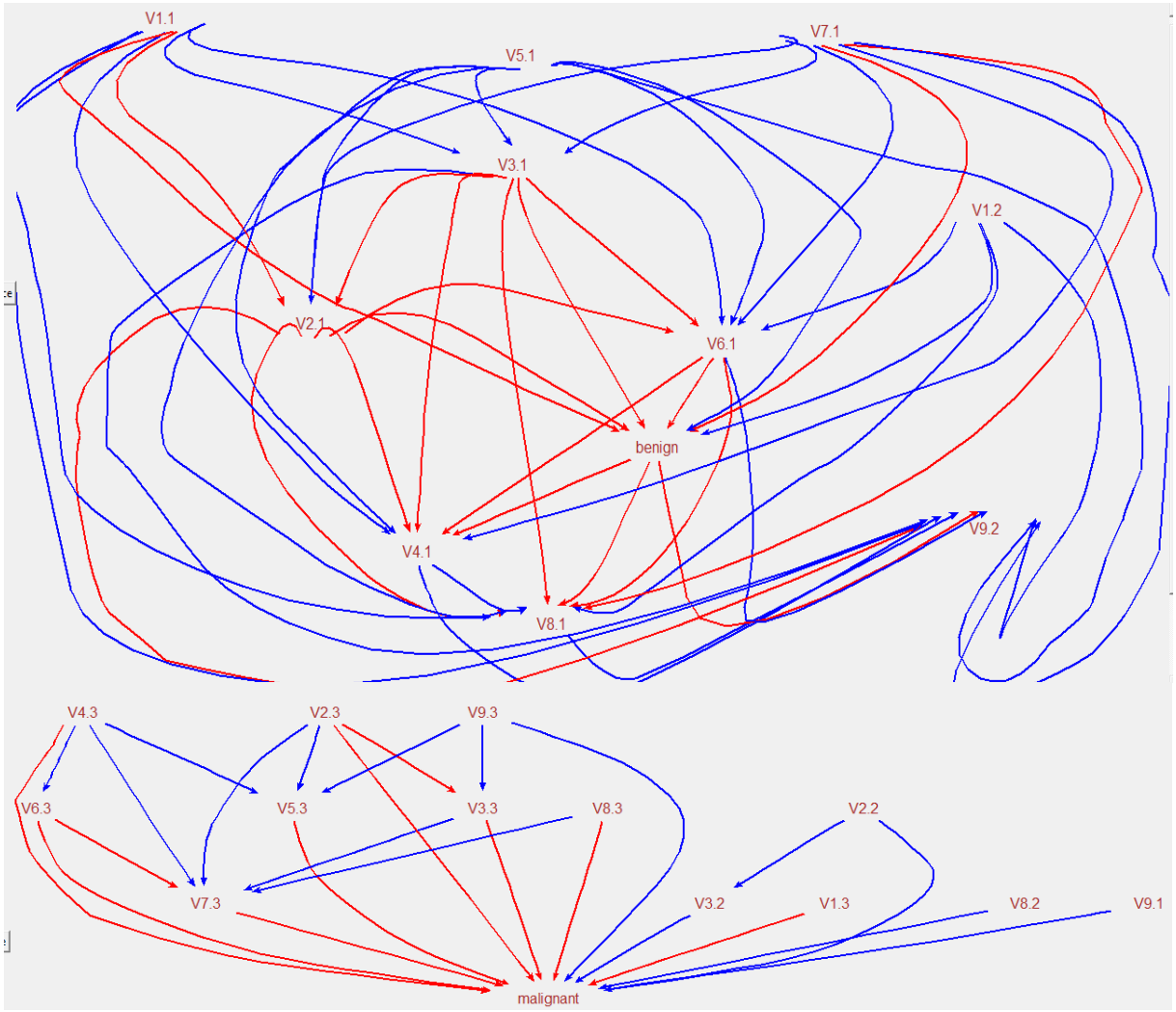

Figure 3.4: graphe implicatif, Seuils : intensité d'implication 70, confiance 70%

3.3.1.2. Application au jeu de données WDBC :

Les figures 3.5, 3.6, 3.7 montrent le graphe implicatif obtenu avec des seuils d'implication de 95, 80,70 et des confiances de 95%,80%,70%.

nous voyons que tout les malades ayant des valeurs fortes dans les variables V13,V11,V26,V8,V4,V3,V24,V21,V23,V27,V6,V24,V7,V8,V14,V11,V29,V30,V25 et des valeurs moyennes dans les variables V14,V11 sont atteintes du cancer du sein WDBC et tout les malades ayant des valeurs faibles dans les variables V1,V3,V27,V28,V23,V21 V2,V22,V5,V9, V25,V16,V30,V18,V29 et des valeurs moyennes dans les variables V19 et V28 et des valeurs fortes dans les variables V15 et V12, ces dernières ne sont pas atteintes du cancer du sein mais les résultats obtenus dans le graphe 3.5 sont les plus fiable car ils ont étaient obtenus avec une confiance de 95%.

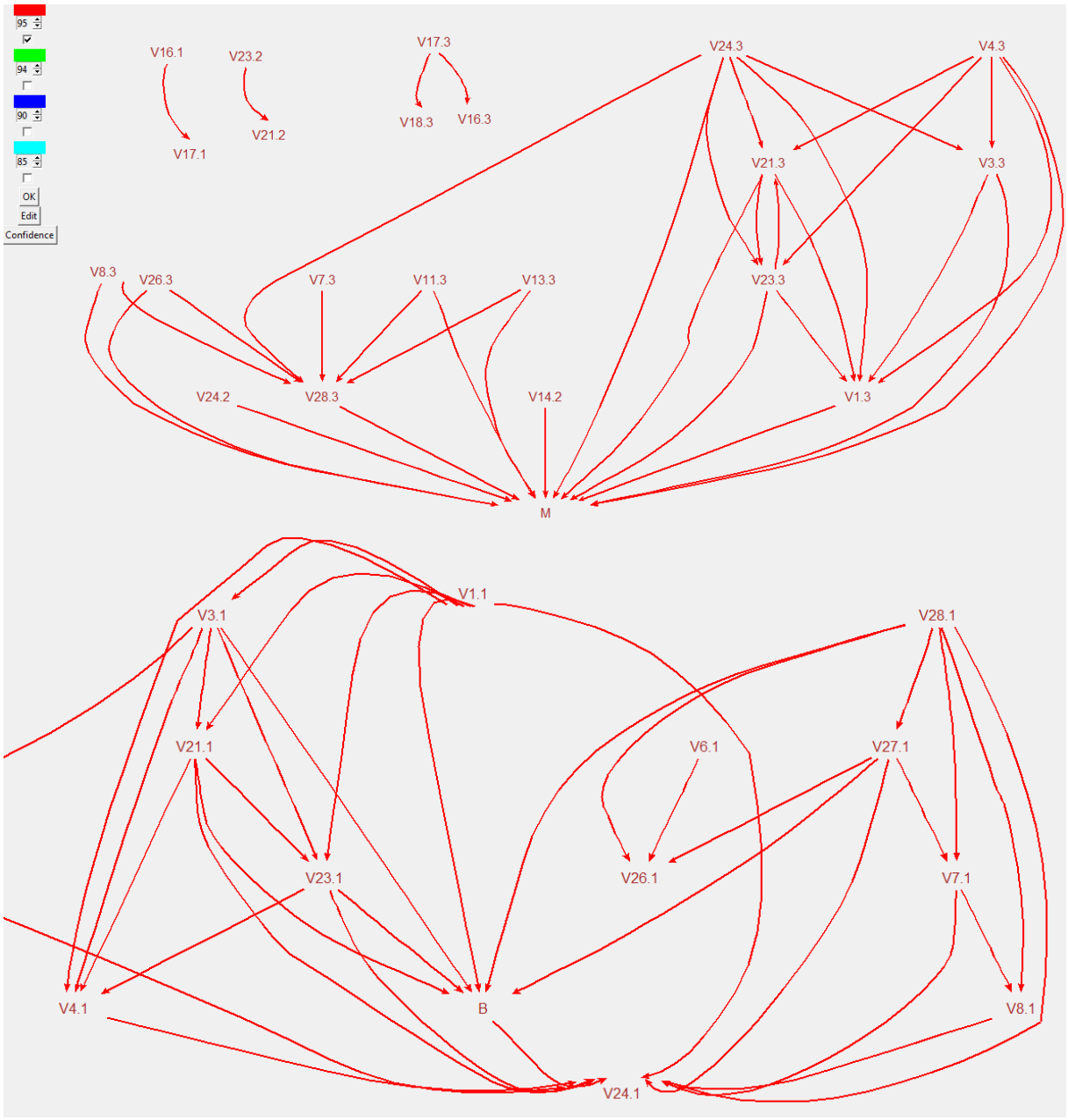

Figure 3.5 : graphe implicatif, Seuils : intensité d'implication 95, confiance 95%

Figure 3.6 : graphe implicatif, Seuils : intensité d'implication 80, confiance 80%

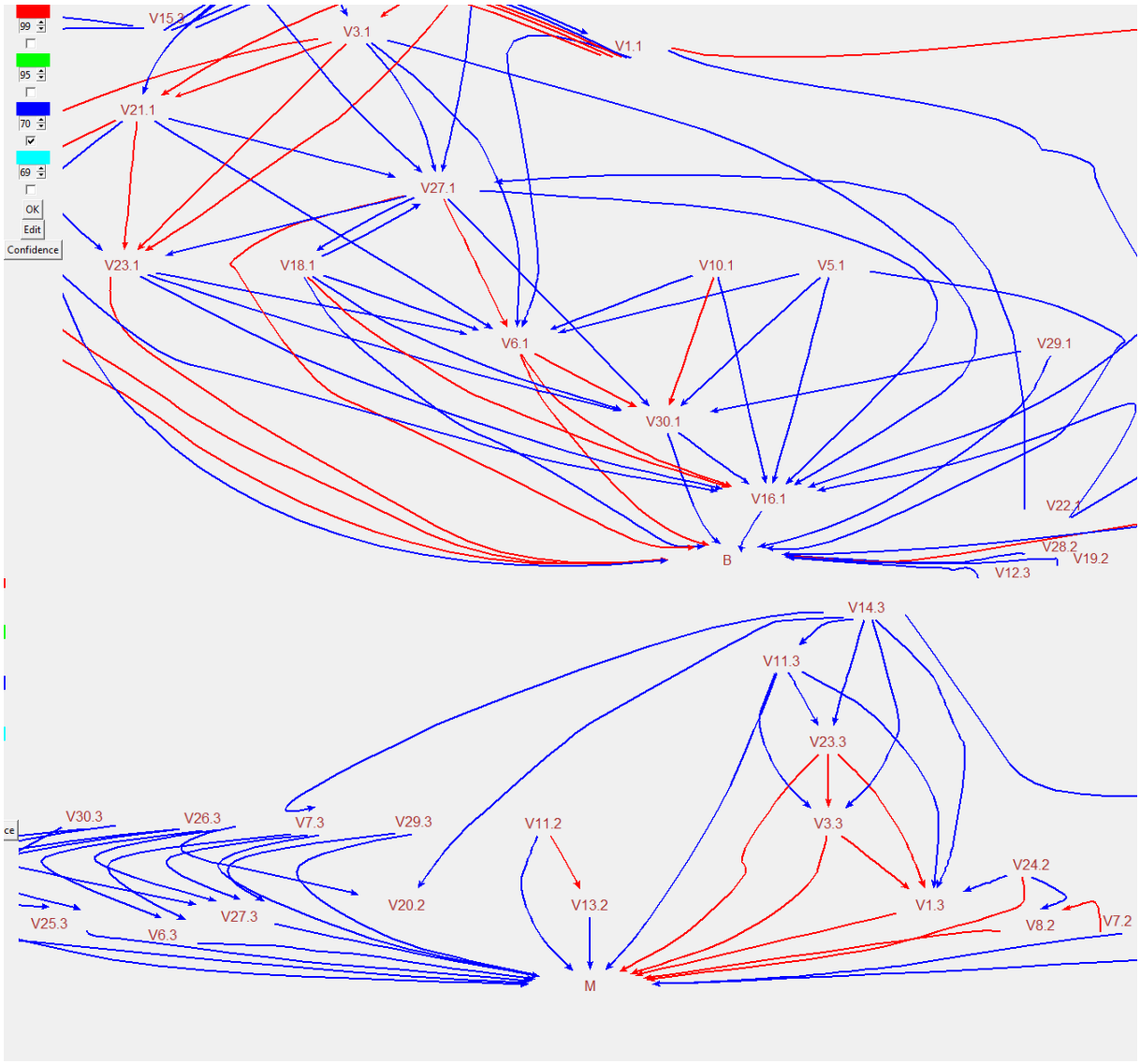

Figure 3.7 : graphe implicatif, Seuils : intensité d'implication 75, confiance 75%

3.3.2. Utilisation de l'implifiance :

Maintenant nous allons appliquer l'implifiance au jeu de données, cette option prend en compte l'implication directe et sa contraposée ainsi que la confiance.

3.3.2.1. Application au jeu de données WBC :

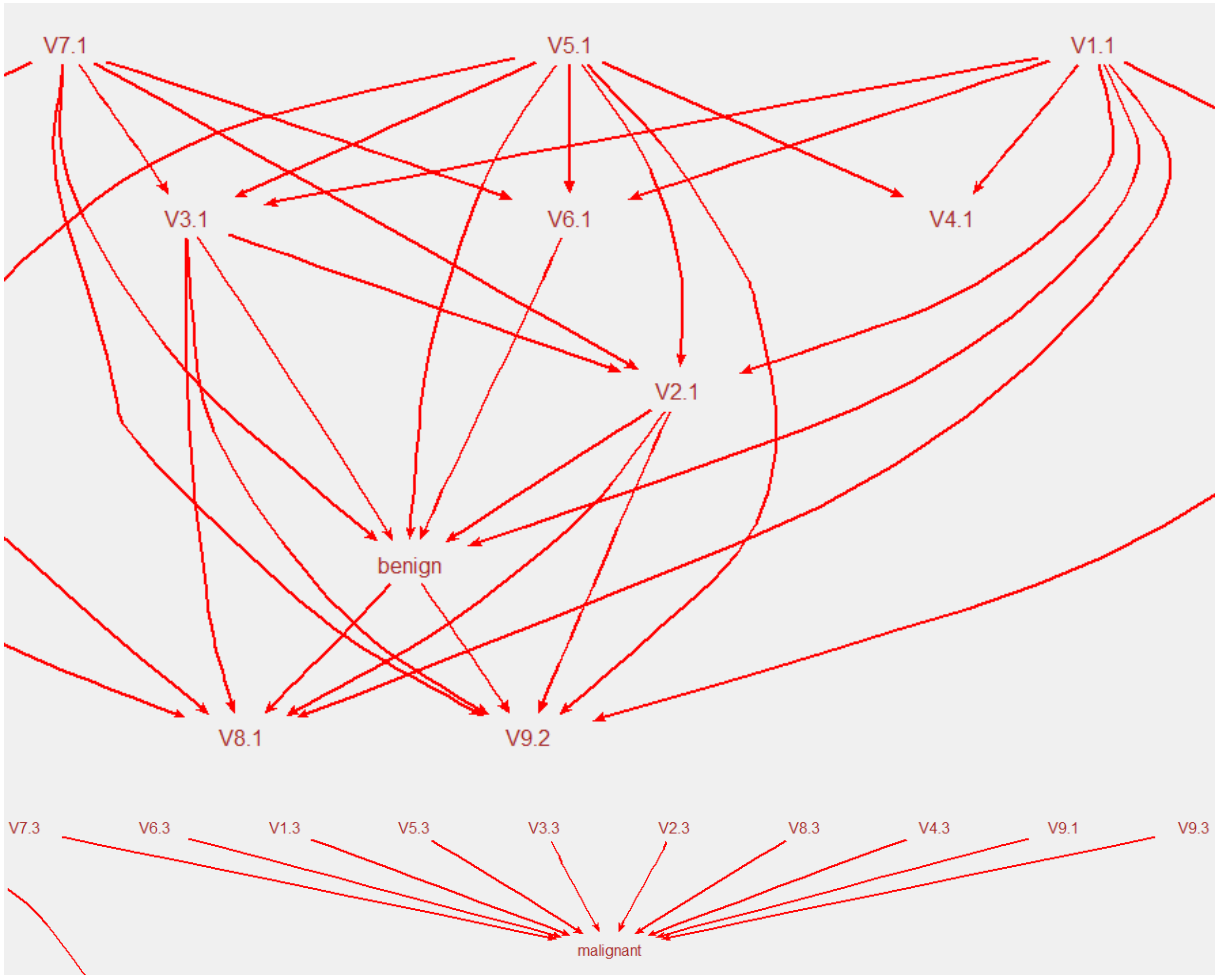

Figure 3.8 : graphe implicatif, seuils 95

Les figures 3.8, 3.9, 3.10 montrent le graphe implicatif obtenu avec des seuils d'implication de 95, 80,70

Nous voyons que tout les cas ayant des valeurs de variable V7.3,V6.3,V1.3,V5.3,V3.3,V2.3,V8.3,V4.3,V9.1,V9.3 avec un seuil d'implication qui égale a 95 plus V2.2 et V4.2 avec un seuil d'implication de 80 plus V3.2,V8.2 avec un seuil d'implication de 70 ces derniers sont tous atteint du cancer.

Figure 3.9 : graphe implicatif, seuils 80

Par contre tout les cas ayant des valeurs de variable de : V3.1, V5.1, V6.1, V2.1, V1.1, V7.1 avec un seuil d'implication de 95 et V7.2, V1.2 avec un seuil d'implication de 80 et 70 ces derniers ne sont pas atteint au cancer

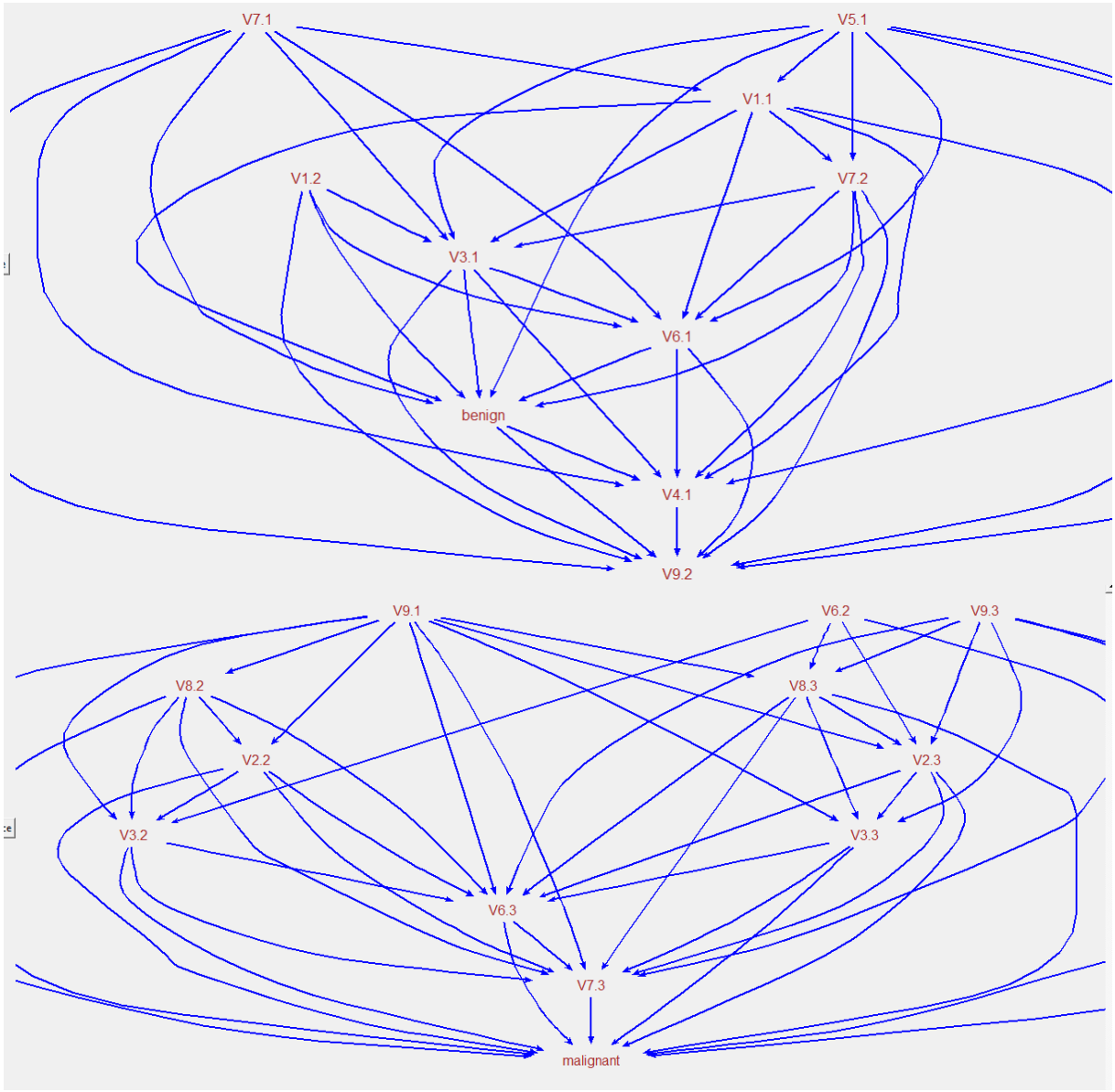

Figure 3.10 : graphe implicatif, seuils 70

3.3.2.2. Application au jeu de données WDBC :

Nous voyons dans les figures 3.11, 3.12, 3.13 que tout les cas qui ont des valeurs dans les variables V23.3, V3.3, V24.2, V28.3, V7.3, V6.3, V8.3, V26.3, V27.3, V29.3, V4.3, V22.3 V26.2, V27.2, V4.2, V11.3, V25.3, V11.2, V13.3, V13.2, V2.3, V7.2, V10.3, V19.3 sont atteint au cancer.

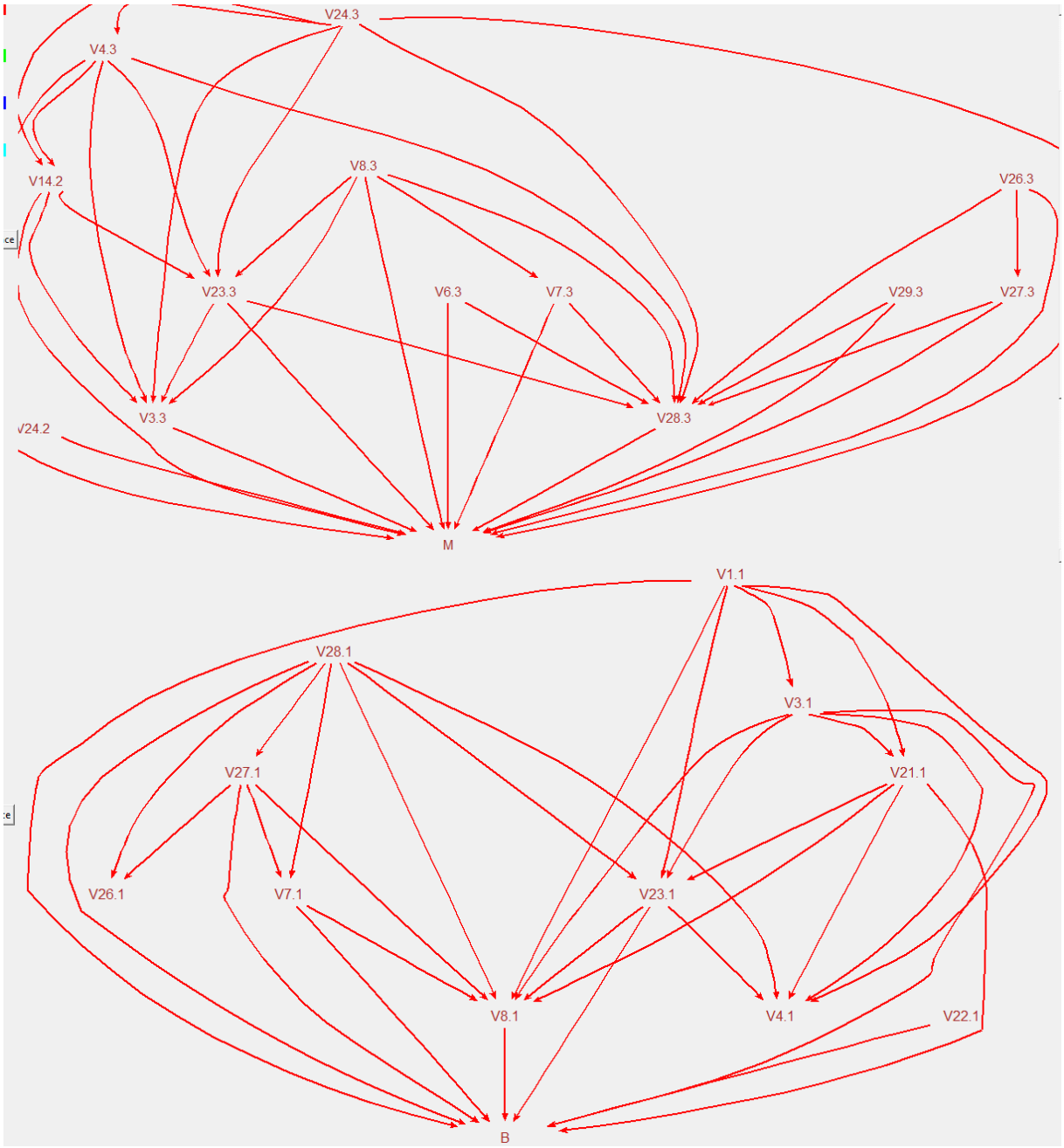

Figure 3.11 : graphe implicatif, seuils 95

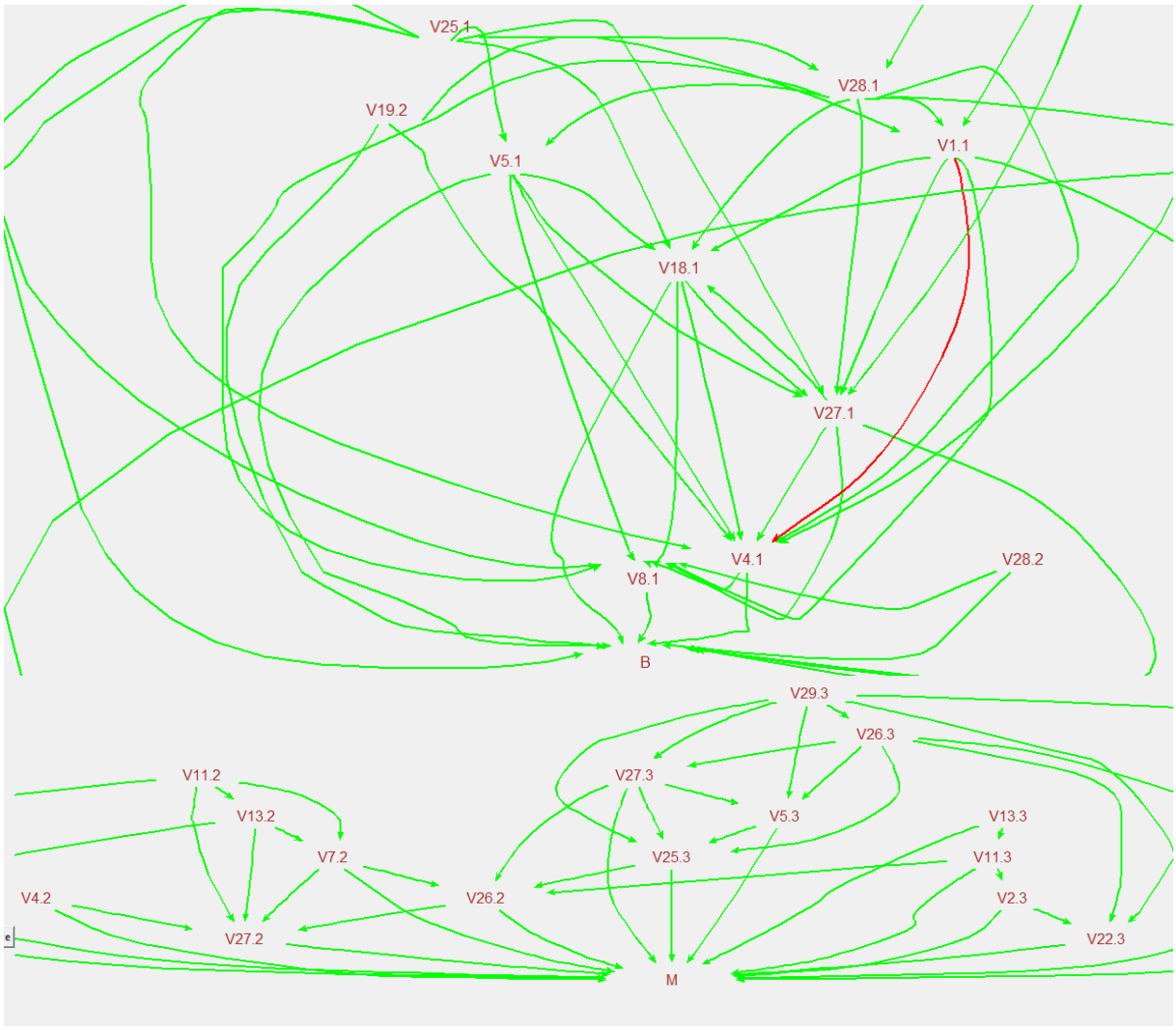

Figure 3.12 : graphe implicatif, seuils 80

Par contre tout les cas ayant des valeurs de variable dans V8.1, V28.1, V7.1, V1.1, V3.1, V21.1, V22.1, V27.1, V28.2, V19.2, V25.1, V5.1, V4.1, V18.1, V30.1, V23.1, V6.1, V2.1, V29.1 ces derniers se sont pas atteint du cancer.

Figure 3.13 : graphe implicatif, seuils 70

3.3.3. Comparaison entre les résultats :

Nous remarquons qu'avec les trois seuils d'implications que presque les mêmes implications dans la conclusion est maline ou bénign (atteint du cancer ou pas) sont obtenus avec les deux méthodes (l'implication classique associé a la confiance et l'implifiance) et avec les deux jeux de données utilisées ,nous avons remarqué que les valeurs fortes des variables conduisent a un cancer du sein ,nous avons aussi remarqué que pour le jeux de données WBC avec les deux options que la variable V9 ne nous donne aucune information sur l'état du patient malgré qu'il existe quelques différences dans les valeurs moyennes mais cela est due au faite qu'avec l'utilisation de l'option d'implication statique associé a la confiance nous avons ajouté l'indice de la confiance ce qui nous a permis d'obtenir juste les implications avec la confiance spécifié mais avec ces résultats nous pouvons constater que les deux options nous conduisent aux même résultats.

3.4. Conclusion :

Dans ce chapitre nous avons appliqué L'ASI a des données qui concerne la maladie du cancer du sein et présenté les résultats dans des graphes implicatifs et nous avons effectué une comparaison entre l'implifiance et l'implication classique associé a la confiance.

Conclusion générale

L’analyse des données permet de traiter un nombre très important de données et de dégager les aspects les plus intéressants de la structure de celles-ci. Le succès de cette discipline dans les dernières années est dû, dans une large mesure, aux représentations graphiques fournies. Ces graphiques peuvent mettre en évidence des relations difficilement saisies par l’analyse directe des données, l'analyse des données est utilisée dans tous les domaines dès lors que les données se présentent en trop grand nombre pour être appréhendées par l'esprit humain. Parmi plusieurs méthodes utilisées nous trouvons l’analyse statistique implicative qui est une méthode non symétrique.

Dans ce mémoire nous avons appliqué l’ASI dans le domaine de la médecine exactement sur des données concernant la maladie du cancer du sein pour savoir les causes de cette maladie. Grace au graphe implicatif nous avons pu obtenir des implications dont la conclusion est malignant ou benign (atteint du cancer ou pas), et puis nous avons comparé entre les deux options disponibles sur Rchic a savoir l’implifiance et l’implication classique associée a la confiance et nous avons déduit que les mêmes implications dans la conclusion est maline ou bénign (atteint du cancer ou pas) sont obtenus avec les deux options.

Comme perspective, nous allons continuer à améliorer ce travail et appliquer l’analyse statistique implicative à d’autres domaines comme la psychologie, la sociologie, la biologie l’économie, bio-informatique, etc. Nous envisageons tout cela dans nos futurs travaux de recherche, qu’ils soient d’ordre individuel, professionnel ou académique.

Bibliographie

[B1] Régis Gras, J.C.Réginer, C.Marinica, F.Guillet, « l'analyse statistique implicative méthode exploratoire et confirmatoire a la recherche de causalité », Cepaduès Ed.Toulouse 201 ISBN :978.2.36493.056.8.(2013).

[B2] GRAS R., KUNTZ P., « Les fondements de l'analyse statistique implicative et leur Prolongement pour la fouille de données », Mathématique et Sciences Humaines, à Paraître, (2001).

[B3] Jean-Claude Régnier, Yahia Slimani, Régis Gras, & Association ARSA «Analyse Statistique Implicative. Des sciences dures aux sciences humaines et sociales », (2015)

[B4] Raphaël COUTURIER, « traitement de l'analyse statistique dans Chic », (janvier 2005).

[B5] Ritschard, Gilbert, Studer, Matthias, Oris, Michel. « Analyse statistique implicative des transitions professionnelles dans la Genève du 19e siècle » In: Gras, Régis and Régnier, Jean-Claude and Marinica, Claudia and Guillet, Fabrice (Ed.). L'analyse statistique implicative. Méthode exploratoire et confirmatoire à la recherche de causa. Toulouse, France : Cépaduès, p. 455-469,(2013).

[B6] Hayette KHALED, Raphael COUTURIER, « apport de la combinaison de la méthode d'analyse statistique implicative (A.S.I.) avec la théorie de réponses aux items (IRT) », (Novembre 2015).

[B7] Jean-Claude Oriol, Anicée Chancel « L'analyse statistique implicative au service d'une étude sur l'anticipation des départs à la retraite », (Novembre 2010)

[B8] Raphaël COUTURIER, « traitement de l'analyse statistique dans Chic », In journées sur l'implication statistique, pages 33-50, Caen, (2000).

[B9] Julien Barnier, Julien Biaudet, François Briatte, Milan Bouchet-Valat, Ewen Gallic, Frédérique Giraud, Joël Gombin, Mayeul Kauffmann, Christophe Lalanne, Joseph Larmarange, Nicolas Robette « introduction a l'analyse d'enquêtes avec R et Rstudio », (Mars 2018).

[B10] Jean-Claude Régnier, Régis Gras, Raphaël Couturier, Antoine Bodin « analyse statistique implicative, points de vue conceptuels, applicatifs et métaphoriques » Université de Bourgogne Franche-Comté Besançon, 614 p. ISBN : 2-9562045-0-5 EAN : 9782956204503, (2017).

[B11] A**.** Bodin, « Analyse implicative : modèles sous-jacents à l'analyse implicative et outils Complémentaires » Prépublication IRMAR No. 97-32, Université de Rennes, (1997).

Webographie

[W1] https://ardm.eu/partenaires/logiciel-danalyse-de-donnees-c-h-i-c/ (consulter en mars 2020).

[W2] https://members.femto-st.fr/raphael-couturier/en/rchic (consulter en mai 2020).

[W3] https://archive.ics.uci.edu/ml/datasets.php (consulter en juin 2020).

Résumé

L'analyse statistique implicative développée par Régis Gras et ses collaborateurs, porte sur les mesures statistiques de qualité des règles d'associations, elle permet d'extraire les règles les plus étonnantes.

L'objectif du présent travail consiste à appliquer l'analyse statistique implicative à des jeux de données en utilisant le logiciel RHIC afin de détecter les causes du cancer de sein a l'aide des graphes implicatifs puis comparer entre l'implifiance et l'implication classique associé a la confiance en utilisant les résultats obtenus avec ces deux options.

Mots-clés : Analyse statistique implicative, RHIC, implifiance, confiance.

Abstract

Implicative statistical analysis developed by Régis and his collaborators relies on quality statistical measures of the association rules, it allows to extract the most astonishing rules.

The purpose of the present work is to apply the implicative statistical analysis to the data sets using the software RHIC to diagnose the causes of breast cancer with the help of implicative graphs then to compare between the implifiance and the classical involvement associated to the confidence using the results obtained with both options.

Keywords: statistical implicative analysis, RHIC, implifiance, confidence.