Chapitre 8 : Graphe de règles d'implication statistique pour le raisonnement courant. Comparaison avec les réseaux bayésiens et les treillis de Galois.

Martine Cadot

Université Henri Poincaré / LORIA, Nancy, France
Martine.Cadot@loria.fr
http://www.loria.fr/~cadot

Résumé. Les règles d'implication statistique ressemblent aux règles du raisonnement mathématique. Ce qui permet de les utiliser facilement pour raisonner sur les données. Toutefois, le modèle sous-jacent aux règles d'implication statistique n'est pas le modèle de la logique formelle utilisé en mathématique, mais un modèle statistique aboutissant à des relations approximatives. Contrairement au raisonnement mathématique, le raisonnement courant se satisfait de règles approximatives. Mais il a besoin d'un graphe pour savoir quels enchaînements de règles sont possibles car en faisant se succéder des approximations, on finit par arriver à des incohérences.
On montrera dans ce chapitre comment fonctionne l'enchaînement de ces règles, notamment à travers la construction du graphe des règles d'implication tel que proposé dans les différentes versions de CHIC et on comparera ce modèle statistique des données à deux autres modèles proches : un modèle algébrique, les treillis de Galois, et un modèle probabiliste, les réseaux bayésiens. Pour permettre des comparaisons aisées, le fonctionnement des trois modèles sera illustré à l'aide d'un même jeu de données médicales librement disponible sur Internet.

1 Introduction

Les règles d'implication statistique ont été conçues à l'origine par Régis Gras dans sa thèse de mathématique (1979). Une règle d'implication statistique entre deux « notions » A et B s'exprime sous la forme « si A alors B » (ou A→B) et indique que si l'élève a acquis la notion A alors il a acquis la notion B. La connaissance de telles règles aide l'enseignant de mathématique à organiser son cours pour faire assimiler un certain nombre de notions à ses élèves. En suivant le modèle de Régis Gras, l'enseignant peut établir un réseau de règles entre des notions (ou plutôt leur assimilation) sans avoir besoin de définir théoriquement ce qu'est chaque notion, du moment qu'il peut déterminer pratiquement si un élève donné l'a acquise ou non. En effet, les règles sont obtenues de façon statistique, par calcul à partir des résultats d'observations d'un ensemble d'élèves, à travers les résultats d'une interrogation écrite par exemple. La méthodologie de construction de règles d'implication statistique n'est pas restée cantonnée à la didactique car elle permet d'établir des liens de « cause à effet »

entre un ensemble quelconque de variables (ou propriétés, items, etc.) à partir de leurs valeurs (présence/absence) pour un ensemble de sujets (ou d'objets, enregistrements, etc.).

Les règles sont obtenues automatiquement à partir de données d'observations, et non postulées une à une avant d'être établies en utilisant un protocole expérimental adapté. Cette façon de procéder s'inscrit dans la fouille de données (Data Mining), qui permet d'explorer après coup des données pas nécessairement recueillies pour cet usage. On parle alors d'extraction de connaissances à partir de données (Knowledge Discovery from Databases). Depuis les travaux fondateurs de Régis Gras, une autre méthode de construction de relations de type causal sur des variables observables a vu le jour, il s'agit des réseaux bayésiens (Naïm 2007), issus des modèles graphiques (Whittaker 1990). Les deux méthodologies de représentation de connaissances que sont l'implication statistique et les réseaux bayésiens ont en commun leur volonté de fournir à l'utilisateur un réseau de règles dans lequel il peut naviguer pour faire des raisonnements. L'utilisateur suit une logique « de bon sens » de type mathématique, c'est pourquoi à ces deux modèles nous en avons ajouté un troisième, plus près, par construction, de la logique mathématique. Il s'agit des treillis de Galois (Godin et al. 1995). La comparaison des trois méthodes, proches par leurs objectifs, mais différentes par leur caractéristiques, va nous permettre de montrer finement les particularités des liens de type causal appelés « règles d'implication statistique ».

Afin de rendre plus concret l'exposé de ce chapitre, nous illustrerons les diverses notions abordées sur un jeu de données médicales, Asia, que nous exposons dans la deuxième partie de ce chapitre. Dans la troisième partie nous donnons les définitions des règles d'implication statistique, nous examinons de façon formelle la façon dont elles vérifient les propriétés attendues par l'utilisateur pour construire un raisonnement, puis plus concrètement le jeu de règles obtenu sur les données Asia. Dans les parties 4 et 5, nous exposons les deux autres modèles de la même façon, et nous comparons les propriétés des jeux de règles de façon formelle ainsi que sur les données Asia. Nous terminons par un bilan et des perspectives. La formulation mathématique dans ce chapitre a été réduite à l'extrême, les définitions nécessitant une écriture mathématique détaillée ont été reportées en annexe, et les preuves formelles ont été mises la plupart du temps en notes de bas de page.

2 Le jeu de données « Asia »

Le jeu de données « Asia » est souvent utilisé pour illustrer le fonctionnement des réseaux bayésiens. Il en existe de nombreuses versions. Nous avons choisi celle fournie par BayesaLab[1] (version d'évaluation, 2001-2008), logiciel de réseau bayésien, qui détaille son utilisation de façon très pédagogique. Le jeu de données initial comportait 10 variables avec des valeurs pour 10 000 sujets. Nous avons retiré les sujets pour lesquels des valeurs manquaient, il en reste 7033, et supprimé la dernière variable (localisation) qui n'était pas utilisée dans le logiciel. Puis nous avons recodé l'âge en deux classes (avant 50 ans et après) au lieu de trois. Ces transformations ont permis un traitement comparable des données par des règles d'implication statistique, des réseaux bayésiens et des treillis de Galois. Dans le tableau 1 figurent les sigles des neuf variables et leur signification, ainsi que le nombre de sujets ayant la caractéristique (valeur 1 pour la variable).

[1] http://www.bayesia.com/fr/produits/bayesialab/release/bayesialab-3-3.php.

Variable	Nombre de valeurs "1"
AG (0 : âge<50 ; 1 : âge>=50)	3391
BR (Bronchite)	3051
CA (Cancer)	271
DY (Dyspnée)	2964
SM (Fumeur)	3271
TC (Tuberculose ou Cancer)	352
TU (Tuberculose)	83
VA (Visite de l'Asie)	74
XR (Rayons X)	685
N (Nombre de sujets)	7033

TAB. 1 – *Liste des variables des données Asia et nombre de leurs valeurs 1.*

3 Les règles d'implication statistique

L'implication statistique, que l'on notera A→B tout au long de ce chapitre, a été définie pour ressembler le plus possible à l'implication logique A⇒B qui est à la base des raisonnements courants. Nous examinons donc les caractéristiques de l'implication logique les plus utilisées pour les raisonnements, et nous regardons si elles se retrouvent à l'identique dans l'implication statistique ou bien si elles sont modifiées et comment tenir compte de ces modifications dans leur utilisation.

Définition de la règle d'implication logique A⇒B. Si A et B sont deux faits qui prennent leurs valeurs dans l'ensemble {V : Vrai, F : Faux}, la règle A⇒B est un fait dont la valeur dépend de celles de A et de B comme indiqué dans le tableau 2. Elle est fausse dans un cas, et vraie dans les trois autres. Pour prouver que la règle logique A⇒B est fausse, il suffit de trouver un « contre-exemple », c'est-à-dire une situation dans laquelle A est vrai et B faux.

	B : Vrai	B : Faux
A : Vrai	A⇒B Vrai	A⇒B Faux
A : Faux	A⇒B Vrai	A⇒B Vrai

TAB. 2 – *Table de vérité de la règle d'implication logique A⇒ B*

Définition de la règle d'implication statistique A→B. Si on connaît les valeurs de N individus pour les faits A et B, la valeur de la règle d'implication A→B se calcule d'après la répartition des N individus en quatre effectifs a, b, c et d selon les valeurs de A et de B, figurant dans le tableau 3. Cette valeur est un nombre compris entre 0 et 1 dépendant des nombres a, b, c et d'autant plus élevé que le nombre b de contre-exemples est petit relativement aux nombres a, c et d.

Graphe de règles d'implication statistique pour le raisonnement courant.

	B : Vrai	B : Faux
A : Vrai	a	b
A : Faux	c	d

TAB. 3 – *Tableau de contingence de A et B pour le calcul de la valeur de la règle d'implication statistique A→B*

Algèbre booléenne ou statistique. Les deux définitions précédentes sont relatives aux mêmes faits A et B ayant pour valeurs Vrai/Faux qui peuvent être codées par 0/1. Mais la première se situe dans une logique booléenne, associée à un univers mathématique abstrait et intemporel, alors que la seconde résulte d'un décompte d'observations plus ou moins objectives faites à un moment donné dans une situation donnée. Il aurait été possible de définir une valeur booléenne pour l'implication statistique en la prenant égale à 1 (Vrai) si l'effectif b est nul (aucun contre-exemple) et à 0 (Faux) sinon. C'est le choix qui a été fait par Guigues et Duquenne (1986) pour définir leurs règles. Nous verrons dans la partie consacrée au treillis de Galois le cadre dans lequel ce choix se situe. Pour la règle d'implication statistique, le choix d'une valeur pouvant être comprise entre 0 et 1 s'appuie sur des considérations statistiques : on compare la valeur de b trouvée pour ces données aux valeurs extrêmes qu'il pourrait atteindre pour des données similaires en cas d'indépendance entre A et B. Le résultat de la comparaison est une probabilité, qui est d'autant plus proche de 1 que b est petit. Pour la calculer, on s'appuie sur des hypothèses concernant les fluctuations possibles des données (loi de Poisson, loi normale, …) comme indiqué dans (Gras et al. 2000). Plusieurs versions de ce calcul existent selon les hypothèses prises sur les lois de probabilité, ainsi qu'une correction utilisant l'entropie mutuelle. Nous précisons maintenant sur un exemple le principe de ce calcul, et nous renvoyons le lecteur intéressé par sa justification théorique à la thèse de Régis Gras ainsi qu'aux articles des différents chercheurs qui ont proposé des corrections (Gras et al. 2001).

Un exemple de calcul de l'indice d'implication statistique. Dans le tableau 4 sont donnés les 4 effectifs des variables Cancer et Bronchite de la base de données Asia afin de les utiliser pour le calcul de la valeur de la règle Cancer → Bronchite.

	Bronchite : Vrai	Bronchite : Faux
Cancer : Vrai	a =139	b = 132
Cancer : Faux	c = 2912	d = 3850

TAB. 4 – *Tableau de contingence de Cancer et Bronchite*

La valeur de b est loin d'être nulle, ce qui pourrait faire penser que la règle est loin d'être vraie. Toutefois la situation de référence n'est plus la règle vraie, avec une valeur théorique de b égale à 0, ni la règle fausse, mais une absence de règle correspondant à l'indépendance statistique entre les deux variables Cancer et Tuberculose. La valeur théorique de b en cas d'indépendance[2] est (b+a)(b+d)/N (pour Asia, N=7033), soit ici 153,4. Quand les données correspondant à une absence de règle fluctuent, leur valeur correspondante de b est supposée suivre une loi normale ayant comme paramètres une espérance E(b) de 153,4 et un écart-type

[2] On suppose ici que le lecteur a des connaissances statistiques concernant l'indépendance de deux variables aléatoires (ou test du Chi2 d'indépendance) et l'utilisation de la loi normale. Dans le cas contraire, il peut retrouver celles-ci dans les ouvrages de statistique de base traitant de l'inférence statistique, comme Morineau, A. (éd.) (1995)..

σ(b) égal à racine(153,4)=12,4. La connaissance de la loi normale nous permet de dire par exemple qu'en théorie 50% des valeurs sont inférieures à l'espérance, 15,9% sont inférieures à l'espérance diminuée de un écart-type (soit 141), 2,3% sont inférieures à l'espérance diminuée de deux écarts-types. La valeur que nous avons trouvée pour b est de 132 soit inférieure à l'espérance de 1,7 écarts-types et, en cas d'absence de règle, on ne pouvait avoir une valeur plus petite que dans 4,2% des cas, c'est-à-dire avec une probabilité[3] de 0,042. Selon la version normale de l'indice d'implication de Régis Gras, la règle Cancer → Bronchite a une valeur de 1-0,042, soit 0,958. Et sa valeur selon la version entropique figurant dans (Gras et al. 2001), dont nous ne donnons pas le détail de calcul ici, est de 0,943. C'est donc une règle qui a de grandes chances d'être vraie.

La règle et sa contraposée ont la même valeur d'indice. Nous venons de voir dans l'exemple que l'indice de A→B se calcule en comparant la valeur de b à la valeur de son espérance E(b)=(b+a)(b+d)/N, les effectifs a, b, c et d correspondant aux valeurs de A et B selon le tableau 3, puis en divisant le tout par la racine carrée de E(b), et que cette valeur de b « centrée réduite » suit la loi normale standard. Pour calculer l'indice de la contraposée de A→B, qui est nonB→nonA, on a écrit les effectifs croisant A et B dans le tableau 5. Ils ont été obtenus à partir de ceux du tableau 3 en échangeant les colonnes entre elles (B devient nonB), les lignes entre elles (A devient nonA) et les lignes avec les colonnes (A devient B et B devient A). Du tableau 3 au tableau 5, les effectifs a et d ont été échangés, alors que b et c sont restés à la même place.

	non A : Vrai	non A : Faux
non B : Vrai	d	b
non B : Faux	c	a

TAB. 5 – *Tableau de contingence de nonB et nonA déduit du tableau 3*

Pour calculer la valeur d'implication statistique de la règle nonB→nonA, on reprend donc la valeur de b que l'on compare comme précédemment à sa valeur théorique en cas d'indépendance E(b)=(b+a)(b+d)/N, et qui ne change pas de valeur quand on échange a et d. Et on divise par la racine carrée de E(b) avant de prendre la probabilité correspondante selon la loi normale standard. Ainsi la règle et sa contraposée ont la même valeur[4]. L'existence de cette propriété facilite l'utilisation des règles d'implication statistique dans les raisonnements.

Les règles A→B et A→nonB ont des valeurs complémentaires. Nous reprenons le tableau 3 en échangeant les colonnes pour obtenir nonB.

[3] Comme b suit la loi normale d'espérance E(b) et d'écart-type s(b), sa valeur centrée réduite $\frac{b - E(b)}{s(b)} = -1{,}73$ suit la loi normale d'espérance 0 et d'écart-type 1, appelée loi normale standard. Selon cette loi, la probabilité d'avoir une valeur inférieure à 61,7 est P(x<-1,73)=0,042.

[4] Nous considérons ici les indices d'implication non corrigés, quelle que soit la loi de probabilité sur laquelle ils s'appuient. Pour les propriétés des indices corrigés, nous invitons le lecteur à consulter (Gras et al. 2001).

Graphe de règles d'implication statistique pour le raisonnement courant.

	non B : Vrai	non B : Faux
A : Vrai	b	a
A : Faux	d	c

TAB. 6 – *Tableau de contingence de A et nonB déduit du tableau 3*

Le calcul de l'indice de A→B nécessite que l'on calcule d'abord la différence entre b et E(b)=(b+a)(b+d)/N. Cette différence est égale[5] à (bc-ad)/N. Le calcul de l'indice de A→nonB s'obtient en remplaçant a par b et c par d et inversement dans cette dernière formule, qui devient (ad-bc)/N, donc les deux différences sont opposées. Puis on divise par la racine carrée de l'espérance, qui est E(b) pour la première règle et E(a) pour la seconde. Cela donne deux valeurs de signe contraire, qui produisent selon la loi normale standard deux probabilités dont l'une est inférieure à 0,5 et l'autre supérieure, la valeur 0,5 étant atteinte en cas d'indépendance entre A et B. Par exemple l'indice d'implication de la règle Cancer→Bronchite vaut 1-P(x<-21,4/racine(153,4))=0,958, et celui de la règle Cancer→non Bronchite vaut 1-P(x<+21,4/racine(117,6))= 0,024. Et l'habitude étant de ne considérer que les règles ayant une valeur bien supérieure à la valeur de 0,5, on n'a jamais simultanément une règle A→B et la règle « contraire » A→nonB. Cette propriété, comme celle du paragraphe précédent, facilite les raisonnements utilisant les règles d'implication statistique.

La règle A→B et sa réciproque B→A. Les effectifs du tableau de contingence lié à la règle B→A se déduisent des effectifs de celui de la règle A→B (cf. tableau 3) par échange des lignes et des colonnes, ce qui produit l'échange de b et c alors que a et d restent inchangés. Les valeurs centrées b-E(b) et c-E(c) sont toutes deux égales à (bc-ad)/N, et les valeurs centrées réduites sont de même signe et égales si et seulement si E(b)=E(c). Ce dernier cas se produit par exemple dès que le « support » de A (le nombre de sujets vérifiant A) est le même que celui de B. Ainsi dès que l'implication statistique de la règle A→B est élevé, celui de sa réciproque l'est aussi quand les supports de A et de B sont proches. L'existence possible de deux règles réciproques l'une de l'autre et de même valeur de vérité ne pose pas de problème en logique formelle, les deux implications étant remplacées par une équivalence. Dans le cas des règles d'implication, si une règle et sa réciproque ont des valeurs élevées, le choix de Régis Gras[6] a été de n'en garder qu'une des deux, celle de plus grande valeur, quand leurs valeurs sont inégales, et de n'autoriser l'équivalence que quand les valeurs sont égales, ce qui arrive assez rarement pour des données réelles avec suffisamment de sujets. Dans le tableau 7 figurent toutes les règles tirées des données Asia ayant un indice d'implication supérieur à 0,95, la valeur de l'indice étant reportée avec cinq chiffres après la virgule. Les treize règles de R1 à R26 ont également leur réciproque d'indice supérieur à 0,95, alors que les 9 règles de R27 à R35 ne sont pas dans ce cas. On remarquera que l'extraction s'est faite parmi les 9 variables de la base de données et parmi leurs négations, soit 18 variables, mais on n'a pas fait figurer les contraposées, sinon il y aurait 70 règles d'indice supérieur à 0,95, chaque règle ayant la même valeur que sa contraposée. Si on choisissait de supprimer, parmi les 2 règles réciproques l'une de l'autre, la règle qui a l'indice le plus petit, on supprimerait les 6 règles R15, R18, R19, R21, R24, R25 et certainement une partie des autres de R1 à R14, si on avait plus de cinq décimales.

[5] Pour établir cette égalité, il suffit de réduire au même dénominateur b et (b+a)(b+d)/N, puis de développer le numérateur après avoir remplacé N par a+b+c+d.

[6] Ce choix a pour but de privilégier un graphe simple, sans boucle ni cycle, permettant des raisonnements de type causal, où un fait ne peut pas être à la fois la cause et la conséquence d'un autre.

Si on fait abstraction des valeurs des coefficients, on peut regrouper les équivalences[7] en 4 classes d'variables :

- Cancer <--> TbOrCa <--> XRay <--> Dyspnea, (R7, R8, R9, R10, R13, R14, R19, R20, R21, R22, R23, R24)
- TbOrCa <--> Tuberculosis <--> XRay (R13, R14, R15, R16, R17, R18)
- Bronchitis <--> Dyspnea <--> Smoker (R3, R4, R5, R6, R11, R12)
- Bronchitis <--> Non Age2 <--> Smoker (R1, R2, R5, R6, R25, R26)

Règles entre deux variables ayant leur réciproque au seuil 0,95			Règles sans leur réciproque au seuil 0,95		
Numéros	Composition en variables	Indices	N°	Composition en variables	Indice
R1;R2	Non Age2 <--> Smoker	1 ; 1	R27	Cancer --> Age2	1
R3;R4	Bronchitis <--> Dyspnea	1 ; 1	R28	TbOrCa --> Age2	1
R5;R6	Bronchitis <--> Smoker	1 ; 1	R29	Cancer --> Smoker	0,99999
R7;R8	Cancer <--> TbOrCa	1 ; 1	R30	Tuberculosis --> Dyspnea	0,99997
R9;R10	Cancer <--> XRay	1 ; 1	R31	TbOrCa --> Smoker	0,99993
R11;R12	Dyspnea <--> Smoker	1 ; 1	R32	XRay --> Age2	0,99971
R13;R14	TbOrCa <--> XRay	1 ; 1	R33	XRay --> Smoker	0,98983
R15;R16	TbOrCa <--> Tuberculosis	0,99999 ;1	R34	Tuberculosis-->nonBronchitis	0,96669
R17;R18	Tuberculosis <--> XRay	1 ; 0,99715	R35	Cancer --> Bronchitis	0,95884
R19;R20	Dyspnea <--> TbOrCa	0,99011 ; 1			
R21;R22	Dyspnea <--> XRay	0.98099 ; 1			
R23;R24	Cancer <--> Dyspnea	1 ; 0,96362			
R25;R26	Non Age2 <--> Bronchitis	0,991 ;0,997			

TAB. 7 – *Liste des règles d'indice d'implication supérieur à 0,95*

Dans la figure 1, on a représenté au sein du même disque les variables qui sont équivalentes entre elles deux à deux.

[7] Deux variables A et B sont liées par une règle d'équivalence, si les règles A→B et B→A ont une valeur à l'indice d'implication supérieure à 0,95. Cette équivalence étant approximative et non exacte, elle n'est pas transitive, comme on le voit dans le paragraphe suivant, ce qui interdit de parler de classes d'équivalence.

Graphe de règles d'implication statistique pour le raisonnement courant.

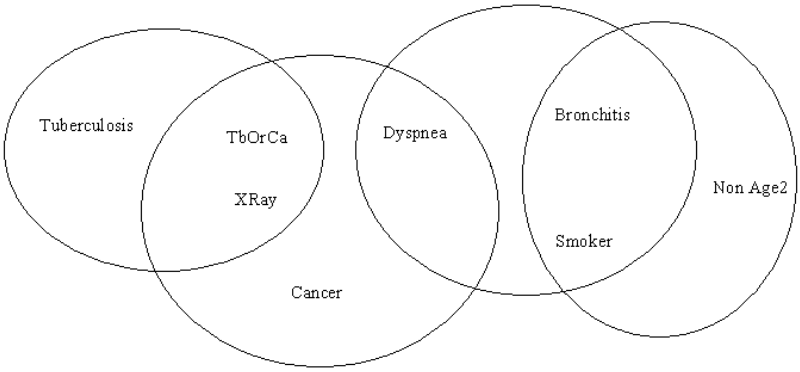

FIG. 1 – *Liste des règles d'indice d'implication supérieur à 0,95*

Les règles A→B et B→C et leur raccourci A→C. En logique formelle, quand les règles A⇒B et B⇒C sont vraies, la règle A⇒C est vraie. Pour le montrer, il suffit de faire la table de vérité croisant A, B et C. L'implication logique n'est fausse que si sa partie gauche est vraie et sa partie droite fausse, et la conjonction (« et ») est fausse dès que l'un des deux est faux. On n'a indiqué dans le tableau 8 que les valeurs fausses pour les 4 dernières colonnes, les valeurs manquantes étant vraies.

On constate que les deux implications A⇒B et B⇒C sont simultanément vraies seulement dans les cas n° 1, 5, 7 et 8, et que la règle A⇒C est vraie dans ces 4 cas. Cette propriété qui associe à deux règles qui se suivent leur raccourci, s'appelle la *transitivité* et fait partie des règles d'un niveau supérieur, les règles d'inférence, qui au lieu d'enchaîner des faits, enchaînent des règles. Elle est un des fondements du raisonnement déductif en logique formelle, mais également en logique courante. Malheureusement, la transitivité des règles (d'équivalence comme d'implication) n'est pas assurée. La figure 1 permet de l'illustrer : si on part de Tuberculosis, situé à gauche, on a la règle Tuberculosis→TbOrCa, TbOrCa→Dyspnea, puis Dyspnea→Bronchitis, mais au lieu d'avoir la règle Tuberculosis→Bronchitis qui en est le raccourci, on a la règle contraire Tuberculosis→non Bronchitis. Pour éviter ce problème qui peut gêner le raisonnement, Régis Gras a choisi de ne garder deux règles qui se suivent que si leur raccourci est de valeur supérieure à 0,5, et de supprimer le raccourci du graphe, dans la mesure où le lecteur le rétablira de lui-même automatiquement s'il a besoin de cette règle.

Cas n°	A	B	C	A⇒B	B⇒C	A⇒B et B⇒C	A⇒C
1	Vrai	Vrai	Vrai				
2	Vrai	Vrai	Faux		Faux	Faux	Faux
3	Vrai	Faux	Vrai	Faux		Faux	
4	Vrai	Faux	Faux	Faux		Faux	Faux
5	Faux	Vrai	Vrai				
6	Faux	Vrai	Faux		Faux	Faux	
7	Faux	Faux	Vrai				
8	Faux	Faux	Faux				

TAB. 8 – *Table de vérité pour 3 variables booléennes*

Pour représenter plus facilement l'ensemble des règles, on utilise l'indice d'implication corrigé[8], on retire les réciproques de moindre valeur et on obtient les 17 règles du tableau 9 ayant la valeur de cet indice supérieure à 0,95, dont il faut retirer les raccourcis pour obtenir la représentation de la figure 2.

R1	Cancer --> TbOrCa	1	R10	Dyspnea --> Bronchitis	0,97592
R2	Tuberculosis --> TbOrCa	1	R11	TbOrCa --> Age2	0,97194
R3	Cancer --> XRay	0,99996	R12	Cancer --> Smoker	0,96957
R4	TbOrCa --> XRay	0,99975	R13	TbOrCa --> Smoker	0,96605
R5	Tuberculosis --> XRay	0,99842	R14	XRay --> Dyspnea	0,96189
R6	Cancer --> Dyspnea	0,97979	R15	XRay --> Age2	0,96095
R7	Cancer --> Age2	0,97972	R16	Tuberculosis --> non Bronchitis	0,95679
R8	TbOrCa --> Dyspnea	0,97937	R17	XRay --> Smoker	0,95453
R9	Tuberculosis --> Dyspnea	0,97856			

TAB. 9– *Les règles d'indice d'implication corrigé supérieur à 0,95, sans réciproques.*

On supprime R3 qui est le raccourci de R1 et R4, ainsi que R5 celui de R2 et R4. On supprime R6 qui est le raccourci de R1 et R8, ainsi que R9 celui de R2 et R8. Bien que l'indice de 0,54 du raccourci de R2 et R11 soit trop petit pour qu'il figure dans le tableau 8 des règles d'indice supérieur à 0.95, sa valeur supérieure à 0.5 lui permet de légitimer la co-existence de ces deux règles R2 et R11. De ce fait, R7, le raccourci de R1 et R11, est supprimé. La règle R10 (avec les règles R2, R8, R16) pose un problème déjà vu précédemment en illustration de la transitivité. On la supprime. Le raccourci des règles R2 et R13 étant d'indice 0.66, la règle R13 est conservée, et la règle R12 est supprimée en tant que raccourci des règles R1 et R11. A ce stade, on a le graphe à gauche de la figure 2. La prise en compte des trois règles R14, R15 et R17 produit la suppression des règles R8, R11 et R13, qui se trouvent être respectivement les raccourcis des règles R14, R15 et R17 avec la règle R4. Il n'y a plus qu'à rajouter la règle R16, et on obtient le graphe à droite de la figure 2.

[8] Cette correction permet de redonner à l'indice d'implication le pouvoir discriminant qu'il a tendance à perdre en cas de données nombreuses, comme c'est le cas des données Asia (voir tableau 7). Toutefois pour l'étude des propriétés de l'indice, nous nous référons à sa version non corrigée.

Graphe de règles d'implication statistique pour le raisonnement courant.

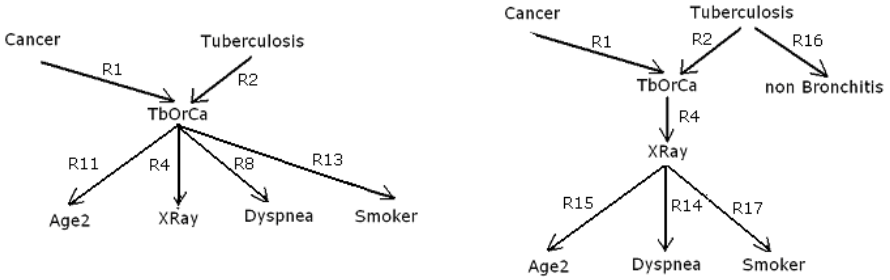

FIG. 2 – *Graphe des implications statistiques (indice corrigé). A gauche une étape intermédiaire de la construction du graphe et à droite le graphe terminé.*

Interprétation et causalité La lecture du schéma de la figure 2 fait apparaître que les implications statistiques, ne peuvent pas s'interpréter de façon directement causale. En effet, ce n'est bien sûr pas le cancer qui cause une augmentation de l'âge des sujets (règle obtenue en suivant les règles R1, R4 et R15), mais éventuellement l'inverse, l'âge avancé qui est une des causes du cancer. Dans le graphe, une bonne moitié des arcs ne peuvent s'interpréter en terme de causalité alors que leurs réciproques pourraient l'être, mais elles ont été éliminées au profit de la règle directe, qui avait une plus grande valeur d'implication. Ce problème d'interprétation n'est pas caractéristique de l'implication statistique : réussir à retrouver les effets et les causes à partir des résultats d'observations faites à un moment donné sans prise en compte d'information supplémentaire est un défi qu'aucune méthode automatique de traitement de données[9] n'a, à notre connaissance, été capable de relever à ce jour.

Pour ce qui est de l'implication statistique, dans sa forme classique, nous avons vu qu'elle est issue du raisonnement mathématique, et qu'elle est équivalente à sa contraposée, ce qui n'est plus le cas dans sa forme entropique (cf. Partie 1). La contraposée de « Cancer → Age2 » est « non Age2 → non Cancer », qui peut se traduire par « un jeune âge a pour effet l'absence de cancer ». Cette traduction en termes de causalité n'est plus du tout choquante car la règle obtenue a une variable explicative, qui est l'âge, en partie gauche et non en partie droite comme précédemment. Ainsi dès qu'une règle d'implication statistique A→B est difficile à interpréter en terme de causalité, alors que sa réciproque B→A serait interprétable, plutôt que de la remplacer par sa réciproque de valeur d'implication inférieure, il est mieux de prendre sa contraposée nonB→nonA qui a la même valeur.

Cas limites : Valeurs implicatives des règles A→A, Faux→A, A→Vrai. Ces règles sont vraies en logique formelle car elles n'admettent aucun contre-exemple. En effet elles ne peuvent avoir leur partie gauche à Vrai et leur partie droite à Faux. Pour calculer leur valeur d'implication statistique, on reprend les valeurs a, b, c et d du tableau 3 qui vérifient b=c=0 pour la première, a=b=0 pour la deuxième, et b=d=0 pour le troisième. L'espérance de b en cas d'indépendance entre partie gauche et droite de la règle est ad/(a+d) dans le premier cas, et 0 dans les deux autres cas. Le calcul de l'indice d'implication statistique est impossible dans ces deux derniers cas, et il donne un résultat très proche de 1 dans le premier cas. Pour les deux derniers cas, on peut choisir 1 si on désire que la règle d'implication statistique

[9] Les sciences expérimentales ont établi des protocoles rigoureux pour essayer d'établir quelles sont les causes parmi un ensemble de causes possibles (cf. exemples dans Cadot 2006).

ressemble le plus possible à la règle logique, 0,5 ou une valeur inférieure si on désire éliminer ces règles sans intérêt statistique. Ce type de calcul peut se produire quand la règle A→B concerne deux variables dont les valeurs sont les mêmes pour tous les sujets (calcul identique à celui de A→A), quand la variable A est fausse pour tous (A→B devient Faux→B) ou quand la valeur de B est vraie pour tous (A→B devient A→Vrai). Ces cas « limites », sous cette forme ou sous une autre équivalente[10] ont fait l'objet de développements dans les articles et livres consacrés à l'implication statistique.

4 Les treillis de Galois

Les *treillis de Galois* sont une représentation algébrique des données d'un tableau booléen de type SujetsXVariables (voir la table à gauche de la figure 5).

1. Définition du contexte
V={a, b, c, d, e}
S={1, 2, …, 7}

	a	b	c	d	e
1	x	x	x	x	x
2		x		x	x
3		x			x
4		x		x	
5	x	x	x		
6	x				
7					

Table de la relation R

2. Calcul de g({d, e})

g({d})={1, 2, 4}
g({e})={1, 2, 3}
donc
g({d, e})={1, 2}

En effet,
g({d, e})=
g({d}∪{e})=
g({d})∩g({e})=
{1, 2, 4}∩{1, 2, 3}=
{1, 2}

3. Calcul de f({1, 2})

f({1}) = {a, b, c, d, e}
f({2}) = {b, d, e}
donc
f({1, 2}) = {b, d, e}

4. Calcul de la fermeture de {d,e}
f•g({d, e})=f{1, 2}={b, d, e}

{d,e} n'est pas fermé, car il est différent de sa fermeture {b, d, e}, rectangle « maximal » grisé dans la table. {b,d,e} est un fermé.

5. Les fermés de 2^V sont ∅, {a}, {b}, {b,e}, {b,d}, {b,d,e}, {a,b,c}, {a,b,c,d,e}
Les fermés de 2^S sont {1,2,…,7}, {1,5,6}, {1,2,…,5}, {1,2,3}, {1,2,4}, {1,2}, {1,5},{1}

FIG. 3 – *Petit exemple de treillis de Galois, détail de calcul d'une fermeture.*

Ces notions, dénommées « treillis de Galois » par Barbut et Monjardet (1970) ou « treillis de concepts » par Wille (1982) sont un des objets d'études de la communauté de recherche « FCA » (Formal Concept Analysis : http://www.upriss.org.uk/fca/fca.html). Nous allons nous contenter ici de décrire la partie du formalisme des treillis de Galois qui concerne la construction des règles logiques entre variables. Puis nous comparerons les règles extraites selon cette vision algébrique des données à l'implication statistique qui provient d'une vision statistique des mêmes données. Cette comparaison se fera dans deux directions : nous examinerons d'abord les propriétés qui découlent de leurs définitions formelles, puis le type d'information qu'elles produisent sur les données Asia.

Définitions. Etant donnés un ensemble de variables V, un ensemble de sujets S, et une relation booléenne R qui les lie, les « treillis de Galois » sont les images des deux *treillis*

[10] Notamment on peut lire dans Gras et al. (2005) que le prolongement par continuité de l'indice d'implication statistique justifie l'attribution de la valeur 0 à la règle A→Vrai.

d'ensemble (voir définition en annexe et dans Davey 1990) associés à V et S par la *correspondance de Galois* associée à R (pour plus de détails voir Mephu Nguifo 1994). Cette « correspondance de Galois » est formée de deux parties duales f et g : f fait correspondre à chaque sous-ensemble de sujets le sous-ensemble de variables qui leur sont liées par R, et g fait correspondre à chaque sous-ensemble de variables le sous-ensemble de sujets qui leur sont liés par R. La figure 3 contient une illustration de ces définitions sur un petit exemple avec 5 variables et 7 sujets avec à gauche le graphe de la relation R (fig 3.1) et à droite un exemple de calcul des correspondances f et g (fig 3.2 et fig 3.3).

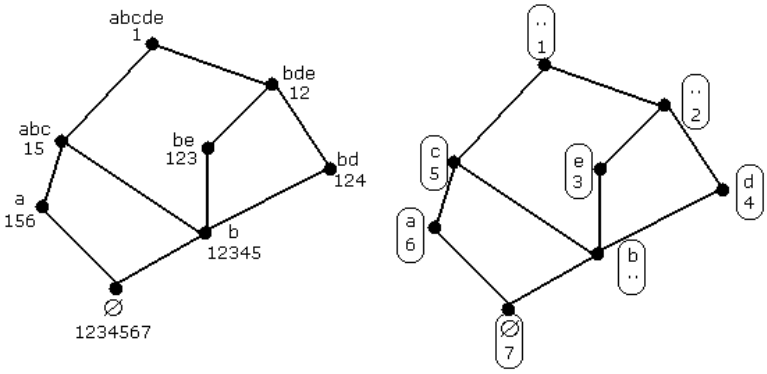

FIG. 4 – *Diagramme de Hasse du treillis de Galois du petit exemple de la figure 3, complet à gauche, simplifié à droite*

La composée des deux est une application qui associe à tout sous-ensemble un sous ensemble le contenant de même nature : de variables si on applique g puis f (fig. 3.4), de sujets si on applique f puis g. Ces deux composées sont appelées des *fermetures*, et les sous-ensembles qui leur sont stables sont appelés des *fermés* (cf. fig. 3.4 et 3.5). La fermeture étant compatible avec l'inclusion, l'intersection et la réunion, les treillis de Galois se trouvent être les deux ensembles de fermés (pour les variables et pour les sujets), munis des opérations ensemblistes et sont isomorphes. Pour retrouver les démonstrations de ces propriétés, on peut se reporter à (Wille 1982). La structure de treillis est représentée par un diagramme de Hasse, qui joint les éléments qui sont ordonnés selon la relation d'ordre, en allant des plus petits (au sens de l'inclusion) aux plus grands, en omettant les raccourcis liés à la transitivité. Comme nous avons deux treillis qui se correspondent, un seul diagramme peut exprimer les deux, chaque élément (ou concept) étant formé d'un ensemble de variables et d'un ensemble de sujets. Afin de faciliter les comparaisons avec l'implication statistique, nous choisissons l'ensemble de variables comme référence et les nœuds sont écrits sur deux lignes, la première relative aux variables, la seconde aux sujets (cf Fig. 4). En bas du diagramme figure le fermé correspondant à l'ensemble de tous les sujets, et en haut celui correspondant à l'ensemble de toutes les variables

Les notions de fermeture et de *règle logique*, c'est-à-dire qui n'admet pas de contre-exemple, sont très liées. En voici une justification rapide (pour plus de détails, voir Guigues et al. 1986) : si un ensemble B de variables, par exemple B={x,y,z,t},contient un ensemble A, par exemple A={y,z}, l'ensemble g(B) est formé des sujets vérifiant toutes les variables

de B, ceux-ci vérifiant nécessairement toutes celles de A, ce qui fait que si A⊂B, g(B)⊂g(A). Cette inclusion est stricte s'il existe des sujets qui vérifient toutes les variables de A sans vérifier toutes celles de B, par exemple qui vérifient y et z mais pas x. Quand ce cas ne se produit pas, A et B ont même fermeture et si on appelle C leur fermeture, par exemple C={x,y,z,t,u,v}, l'inclusion g(C)⊂g(B)⊂g(A) est en fait une égalité. Tous les sujets vérifiant les variables de A vérifient donc aussi celles de B et de B-A (variables de B ne figurant pas dans A), c'est-à-dire qu'il n'y a pas de contre-exemple à la règle A→B-A, qui s'écrit ici {y,z}→{x,t}, ou plus simplement yz→xt. Nous l'appelons règle logique pour la différencier de la règle d'implication statistique.

Nous venons de définir une règle logique à l'aide des deux ensembles emboîtés A et B de même fermeture C. Il peut y en avoir d'autres définies sur le même ensemble de sujets, ce sont par exemple les règles yz→x, yz→t, yz→u, yz→v, yz→uv, xuz→u, yzt→v, etc, s'écrivant de façon plus générale A∪D→E, où D et E sont deux parties disjointes[11] de l'ensemble C-A contenant les quatre variables x, t, u et v. La partie gauche de ces règles est un ensemble non fermé, appelé *lacune* par Guigues et Duquenne (1986). Ces auteurs ont proposé un parcours du treillis de Galois permettant d'extraire seulement une partie génératrice des règles logiques : pour chaque ensemble fermé C de variables, ils cherchent les lacunes A dont C est la fermeture et créent, si certaines conditions sont vérifiées, les règles correspondantes sous la forme A→C-A, ce qui donne par exemple la règle yz→xtuv. Pour chacune de ces règles, on peut déduire les autres règles construites sur les mêmes sujets en retirant des variables du membre droit de la règle, et en en mettant éventuellement une partie dans le membre gauche. Nous avons déjà évoqué dans la partie précédente au sujet de la transitivité ce procédé, appelé *règle d'inférence* qui, appliqué à des règles d'un ensemble, permet d'en créer de nouvelles. Parmi les règles d'inférence proposées par Guigues et Duquenne, il y a, outre la transitivité, deux règles qui permettent de modifier le nombre de variables des membres gauches et droits des règles, une comme celle que nous venons de voir à partir d'une seule règle, et l'autre à partir de deux. Si on applique l'algorithme proposé par Guigues et Duquenne au cas du petit exemple de la figure 3, on obtient les 6 règles e→b, d→b, c→ab, ab→c, ae→d, cd→e, qui suffisent[12] pour engendrer au moyen des règles d'inférence les 72 règles logiques associées à la relation R.

On ne peut pas lire directement les règles sur le diagramme de Hasse mais il est possible de les obtenir à partir d'une simplification de ce diagramme obtenue en retirant pour chaque nœud les variables figurant dans les nœuds en dessous et les sujets figurant dans les nœuds au dessus, comme dessiné à droite dans la figure 4. Si une arête relie deux nœuds contenant chacun une variable, la règle correspondante s'écrit en parcourant l'arête du haut vers le bas : pour la figure 4, on obtient c→a, c→b, e→b et d→b. On peut retrouver les règles plus complexes en remontant dans le diagramme à partir de deux nœuds (partie gauche de la

[11] On choisit, en général, une partie E non vide, bien que la règle {y,z}→∅ soit cohérente avec le formalisme algébrique.

[12] Les règles selon le formalisme de Guigues et Duquenne s'écrivent en mettant en partie droite la réunion des parties gauche et droite habituelles, ce qui donne e→be, d→bd, c→abc, ab→abc, ae→ade, cd→cde. Les trois règles d'inférence sont mr1, qui est la transitivité, mr2, qui transforme une règle par adjonction de variables quelconques, les mêmes en partie gauche et droite, et mr3, qui associe à deux règles la règle formée de la réunion des deux parties gauches et droites. Il est à noter qu'il peut arriver que certaines règles générées ne soient vérifiées par aucun sujet. Ce n'est pas le cas des 72 règles de ce petit exemple qui sont toutes vérifiées par au moins un sujet : le sujet 1.

règle) jusqu'à leur maximum en suivant les arêtes, puis en redescendant dans un certain nombre de nœuds en suivant les arêtes à partir de ce maximum (partie droite de la règle). Ainsi dans la figure 4, si on remonte à partir de a et b, on arrive à c ce qui donne la règle ab→c, et si on part de c et d, on doit remonter jusqu'en haut, ce qui fait qu'on peut alors redescendre dans chaque nœud et on obtient par exemple la règle cd→abe.

Réciproques et équivalences. Nous avons vu dans le paragraphe précédent de définitions du treillis de Galois que le diagramme de Hasse simplifié (cf. à droite de la figure 4) se lit plus souvent du haut vers le bas qu'inversement pour trouver les règles logiques. Du fait de ces deux sens possibles de lecture, il découle qu'un jeu de règles peut contenir une règle et sa réciproque. Par exemple les deux règles c→ab et ab→c ont été trouvées à partir du petit exemple de la figure 3. Plus généralement, les règles A→B et B→A coexistent si et seulement si les deux ensembles disjoints de variables A et B sont vérifiés par le même ensemble de sujets. Il peut arriver que ces deux ensembles de variables forment un seul nœud, ce qui rajoute aux deux façons que nous avons exposées de lire des règles à partir du diagramme de Hasse simplifié une troisième façon : boucler (c'est-à-dire joindre un nœud à lui-même) sur les nœuds qui contiennent plus d'une variable.

Règles non informatives. Dans le cas où l'ensemble de variables B est inclus dans l'ensemble A, les sujets qui vérifient A vérifient également B, ce qui permet d'écrire la règle A→B, par exemple xy→x. Cette règle est appelée règle « non informative » par Guigues et Duquenne (1986). La règle A→A fait également partie de ces règles non informatives. D'autres règles n'apportent pas non plus d'information, comme la règle A→B, où A est un ensemble quelconque de variables, et B un ensemble de variables vérifiées par tous les sujets (notamment, B peut se réduire à l'ensemble vide), ainsi que la règle A→B, où A est un ensemble de variables vérifié par aucun sujet et B un ensemble quelconque de variables.

Variables et négation. La structure de treillis de Galois n'est pas compatible avec la négation des variables. Notamment le treillis des négations peut comporter un nombre de nœuds différent (cf. figure 5). Tant que les règles du treillis des variables ont une variable en partie gauche et une en partie droite, on retrouve leurs contraposées dans le treillis des négations. Ce n'est pas toujours le cas des règles plus complexes, comme ab→c qui se réécrit « a nonc→non b » et « b non c →non a », aucune de ses règles, composées en partie seulement de négations de variables, ne figure dans le treillis des négations.

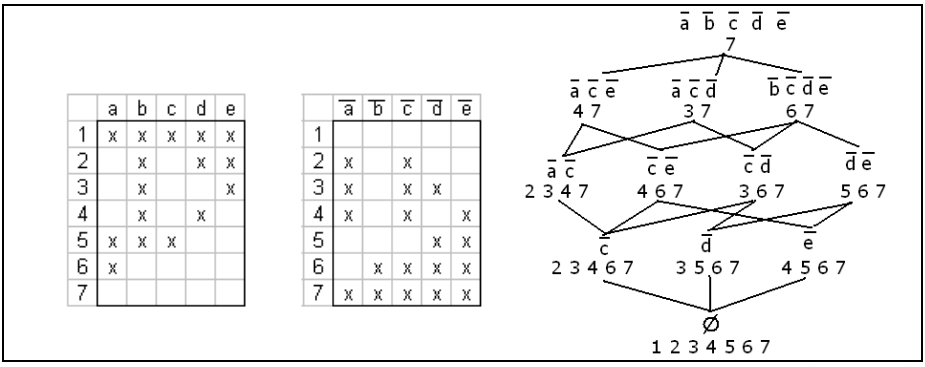

FIG. 5 – *A gauche, variables du petit exemple de la figure 2, au milieu leurs négations, à droite le diagramme de Hasse des négations.*

Application aux données Asia. Nous avons construit le treillis de Galois des données qui comporte 140 fermés et l'algorithme de Guigues et Duquenne de recherche d'une partie génératrice de règles nous a fourni 17 règles. Les voici, avec les noms de variables sous forme de deux lettres (la correspondance entre ces deux lettres et l'intitulé est faite dans le tableau 1), suivies de leur support, c'est-à-dire du nombre de sujets qui vérifient leurs parties gauche et droite, entre parenthèses :
TU→TC (83), CA→TC (271), TC VA→DY XR (7), DY XR VA→ TC (7), DY SM VA→ BR (13), BR CA→XR (139), AG SM TU→XR (14), CA TU→AG XR (2), AG BR SM TC→XR (87), AG BR SM VA→DY (6), CA VA→AG (1), BR VA XR→DY SM TU (1) et AG BR SM TU→DY XR (7), BR CA TU→DY SM VA (0), CA TU VA→BR SM (0), CA DY SM TU→BR VA (1) et AG BR TC VA→CA (0)

Notons que parmi ces règles, quatre ont un support nul, qui proviennent du fermé correspondant à toutes les variables et à aucun sujet, élément maximal du treillis de Galois, nœud au sommet du diagramme de Hasse. Bien qu'elles paraissent avoir peu d'intérêt, n'étant vérifiées par aucun sujet, elles sont correctes du point de vue algébrique, dans la mesure où elles n'admettent aucun contre-exemple. Avec les règles d'inférence, elles produisent des règles de support nul également. En combinant les 17 règles au moyen des règles d'inférence, on retrouve les 2992 règles exactes présentes dans les données Asia dont 514 de support non nul.

Comparaison des deux jeux de règles (implication statistique et règles logiques) extraits des données Asia. Les deux jeux de règles ont en commun seulement deux règles, qui sont TU (Tuberculosis) → TC (TbOrCa) ; CA (Cancer) → TC (TbOrCa). Parmi toutes les règles d'implication statistique, ce sont les seules qui ont un indice d'implication statistique corrigé égal à 1 (voir tableau 9). Et parmi toutes les règles logiques, ce sont les seules possédant une seule variable en partie gauche, et elles font partie des 17 règles génératrices. Pour faciliter la comparaison des règles qui ne sont pas communes aux deux jeux, on se limite à une partie représentative de ces règles pour chacun des deux modèles. Dans le dernier, on ne considèrera que les 13 règles logiques de support non nul de la partie génératrice, et dans le premier que les 7 règles de la figure 2.

Graphe de règles d'implication statistique pour le raisonnement courant.

On peut remarquer dans la figure 2 que la variable VA (Visit in Asia) a disparu du graphe d'implication statistique alors qu'elle est présente dans la partie gauche de presque la moitié des règles logiques de la partie génératrice. Parmi les 7033 sujets de la base de données, seuls 74 ont une valeur de 1 à cette variable, donc une partie des règles logiques a été construite sur environ 1% des données. Cette variable est présente dans la partie gauche des 6 règles TC VA→DY XR (7), DY XR VA→ TC (7), DY SM VA→ BR (13), AG BR SM VA→DY (6), CA VA→AG (1), BR VA XR→DY SM TU (1) de supports compris entre 1 et 13. Cela montre la différence de point de vue qui existe entre la représentation des données par un treillis de Galois et celle par un graphe d'implication statistique : dans le premier cas on privilégie l'extraction des règles qui caractérisent de façon exacte des groupes de sujets, quelle que soit la taille de ces groupes, même réduits à une personne[13], dans le second on cherche des relations suffisamment « fortes » (avec un fort indice d'implication statistique) entre variables pour que quelques sujets particuliers les contredisant ne puissent pas les remettre en cause.

Les règles des deux types, implication statistique et règle logique, sont créées à partir des mêmes comptages de sujets, et il en résulte une certaine ressemblance dans leur orientation (membre gauche et droit). Dans la figure 2, on peut remarquer qu'une variable comme Age (resp. Smoker, Dyspnea, non bronchitis) vérifiée par un nombre de sujets égal à 3391 (resp. 3271, 2964, 3622), est en nœud terminal, donc dans le membre droit des règles, alors qu'une variable comme Cancer (resp. TbOrCa, Tuberculosis, Xray) qui a un nombre de sujets égal à 271 (resp. 352, 83, 685) est dans un nœud non terminal, et apparaît dans le membre gauche d'une plus grande proportion de règles. On retrouve ce phénomène dans les règles logiques issues du treillis de Galois : nous avons décrit précédemment le parcours du diagramme de Hasse permettant de retrouver les règles logiques (cf. figure 4), dont le départ est généralement dans des nœuds situés plus haut que ceux de l'arrivée. Comme les nœuds les plus bas correspondent à plus de sujets, on obtient ainsi une plus grande proportion de règles logiques avec des variables fréquentes en membre droit qu'en membre gauche.

5 Les réseaux bayésiens

La création de ce type de réseau répond, d'après Olfa Ben Naceur-Mourali et Christophe Gonzales (2004) à un besoin d'assouplissement des systèmes experts à base de règles, ces derniers ne fonctionnant qu'avec des faits certains. D'après les auteurs, MYCIN, créé en 1976 par Shortliffe, fut une première tentative d'introduction de l'incertitude au moyen de "facteurs de certitude", mais la première formalisation véritablement opérationnelle de cette incertitude par des probabilités date de Pearl (1988) : ce sont les réseaux bayésiens, qui sont une représentation par un graphe de liens probabilistes entre faits.

Reprenons l'exemple très pédagogique que donne Jensen dans son introduction aux réseaux bayésiens (Jensen 1996) en nous autorisant toutefois une certaine liberté d'adaptation de l'histoire qu'il raconte :

« Sherlock Holmes sort de chez lui le matin pour aller travailler. En arrivant à sa voiture, il constate que sa pelouse est mouillée alors qu'il ne pleut pas. Il se dit qu'il a dû oublier de

[13] Dans le formalisme de treillis de Galois, il est même licite d'envisager un groupe particulier de sujets, le groupe « vide », qui vérifie toutes les règles de support nul.

couper son système d'arrosage la veille au soir. Il se dirige vers la cave pour aller l'arrêter quand il jette un coup d'oeil sur la pelouse de son voisin Watson : elle est mouillée. Il rebrousse alors chemin et monte dans sa voiture pour se rendre à son travail. »

Nous allons procéder en cinq étapes pour montrer la modélisation de ce raisonnement par un réseau bayésien : 1) définir un réseau de faits et de règles sur lequel le raisonnement peut s'appuyer, 2) lui ajouter les probabilités conditionnelles pour en faire un réseau bayésien, et montrer le raisonnement « direct » de Holmes aboutissant à sa première décision 3) Présenter le moteur du réseau bayésien, formé d'hypothèses et de formules qui ne seront pas remises en cause 4) montrer le raisonnement « bayésien » de Holmes aboutissant à la deuxième décision et 5) à la dernière décision.

Raisonnement causal avec quatre faits et trois règles. Son raisonnement peut être représenté comme un parcours dans un réseau causal comportant quatre faits et trois relations, voici les quatre faits qui peuvent prendre la valeur "Vrai" ou "Faux" :

- H : La pelouse de Holmes est mouillée
- W : La pelouse de Watson est mouillée
- A : Le système d'arrosage de Holmes n'a pas été coupé
- P : Il a plu

Et voici les trois relations causales qui les lient :

- P→H : s'il a plu, la pelouse de Holmes est mouillée
- P→W : s'il a plu, la pelouse de Watson est mouillée
- A→H : si le système d'arrosage de Holmes n'a pas été coupé alors la pelouse de Holmes est mouillée

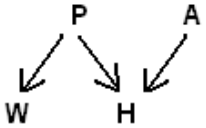

FIG. 6 – *Petit exemple de réseau bayésien emprunté à Jensen (1995)*

Holmes utilise l'information au fur et à mesure qu'elle arrive pour évaluer la probabilité qu'il ait oublié de couper l'arrosage. Ainsi, il prend trois décisions successives contradictoires:

- Le risque que A soit vrai est a priori faible. Il va directement à sa voiture
- Il constate que H est vrai. Cela augmente suffisamment à ses yeux le risque que A soit vrai pour justifier un détour par la cave
- Il constate que W est vrai. Cela diminue suffisamment à ses yeux le risque que A soit vrai pour ne plus justifier un détour par la cave. Il rebrousse chemin et monte dans sa voiture.

Les probabilités fixées au départ. Pour modéliser son comportement, il faut ajouter des probabilités. On associe à chaque fait ayant plusieurs causes une table de probabilités donnant toutes les probabilités de ce fait conditionnellement aux autres. Les probabilités proposées par l'auteur pour ces faits sont dans le tableau 10.

Graphe de règles d'implication statistique pour le raisonnement courant.

P=1	0.2
P=0	0.8

P(P)

A=1	0.1
A=0	0.9

P(A)

	P=1	P=0
W=1	1	0.2
W=0	0	0.8

P(W|P)

	P=1		P=0	
	A=1	A=0	A=1	A=0
H=1	1	1	0.9	0
H=0	0	0	0.1	1

P(H|P,A)

TAB. 10 – *Les probabilités conditionnelles des 4 faits de la figure 6*

Explicitons ces tableaux. Celui de gauche n'a que deux lignes car aucun arc n'arrive à P (cf. figure 6}, ce qui signifie qu'il n'est l'effet d'aucune cause dans ce modèle. On a en première ligne la probabilité 0.2 que P soit vrai (P(P=1)=0.2 en notant Vrai par 1 et Faux par 0), donc qu'il ait plu cette nuit. Et dans la ligne suivante 0.8, qui est la probabilité qu'il n'ait pas plu. Pour toutes les colonnes de ces tableaux la somme sera 1. A côté on a le même type de tableau pour A. Notons que Holmes ne fait pas de détour pas la cave car il estime la valeur de P(A=1) de 0.1 trop faible pour prendre la peine de ce détour. Le tableau de P(W|P) contient 4 valeurs. En haut à gauche, c'est la probabilité que W soit vraie, sachant que P est vraie. Elle est de 1, ce qui signifie que s'il a plu, on a la certitude que la pelouse de Watson est mouillée. Par contre s'il n'a pas plu, il y a quand même une probabilité de 0.2 pour que la pelouse de Watson soit mouillée, certainement parce qu'il l'arrose parfois lui aussi. Le dernier tableau s'interprète de la même façon en fonction des deux causes que sont A et P. Que Holmes ait oublié ou non le système d'arrosage ouvert (A=1 ou 0), s'il a plu, on est certain que sa pelouse est mouillée. Par contre s'il n'a pas plu, la probabilité pour que la pelouse soit mouillée est nulle s'il n'a pas arrosé et de 0.9 dans le cas contraire.

Le « moteur » probabiliste du raisonnement. Pour pouvoir faire fonctionner le modèle, il faut calculer les probabilités conjointes[14], marginales.

P=1	0.2
P=0	0.8

P(P)

A=1	**0.1**
A=0	0.9

P(A)

	P=1	P=0	
W=1	0.2	0.16	0.36
W=0	0	0.64	0.64

P(W,P) P(W)

	P=1		P=0		
	A=1	A=0	A=1	A=0	
H=1	0.02	0.18	0.07	0	0.27
H=0	0	0.1	0.01	0.72	0.73

P(H,P,A) P(H)

TAB. 11 – *Les 4 probabilités marginales et les 2 conjointes déduites du tableau10*

Les valeurs des probabilités des lois conjointes sont calculées d'après les formules de la note de bas de page 10, et celles des deux lois marginales P(W) et P(H) sont obtenues en sommant les cellules des lignes des lois conjointes. Ces formules sont le « moteur » du réseau bayésien. Elles ne changeront pas pendant son fonctionnement. Par contre les probabilités de P (il a plu) et de A (le système d'arrosage était en marche) sont des probabilités a priori, susceptibles de changement. C'est d'ailleurs un des buts du réseau bayésien que d'aider à les mettre à jour en fonction de nouvelles informations, et d'en déduire leurs probabilités a posteriori.

Prise en compte de l'information « H vrai ». On remonte l'*arc* AH (ce qu'on appelle arc est un lien orienté de A vers B correspondant à la règle A→H) de la conséquence H à la

[14]Loi conjointe de P et W : P(W,P)=P(W|P)P(P), Loi conjointe de A, P et H : P(H,P,A)=P(H|P,A)P(P,A), et avec A et P indépendants, P(P,A)=P(P)P(A).

cause A en utilisant la formule de Bayes (aussi appelée de *probabilités des causes*)[15]. Pratiquement cela consiste à multiplier les probabilités conjointes de P(H,P,A) par un coefficient pour que la première ligne ait pour somme 1 (c'est le quotient de la nouvelle valeur sur l'ancienne de P(H=1)), et pareillement pour que la deuxième soit de somme nulle. Une fois les valeurs de P(H,P,A) réactualisées, on réactualise aussi celles de P(A) et P(P) par sommation et on reprend le sens "normal" de l'arc PW pour réactualiser P(W,P), ce qui donne les valeurs du tableau 12.

					P=1	P=0				P=1		P=0			
										A=1	A=0	A=1	A=0		
P=1	0.74	A=1	**0.34**	W=1	0.74	0.05	0.79	H=1	0.07	0.66	0.26	0	1
P=0	0.26	A=0	0.66	W=0	0	0.21	0.21	H=0	0	0	0	0	0
P(P)		P(A)		P(W,P)			P(W)	P(H,P,A)					P(H)

TAB. 12 – *Modification du tableau 11 prenant en compte la certitude que H est vrai*

On constate à la lecture de ce tableau que P(A=1) est passé à 0.34, alors qu'il était de 0.10 dans le précédent. Ainsi la prise en compte de l'information sur H a augmenté la probabilité que A soit vrai, ce qui a poussé Holmes à faire un détour par sa cave.

Prise en compte de l'information supplémentaire « W vrai ». On met à jour les valeurs de la même façon que précédemment (cf. tableau 13) en commençant par la loi conjointe de P(W,P), en réactualisant celles de P(P) et P(A) par sommation, et en terminant pas celle de P(H,P,A). La probabilité de A[16] redescend à 0.16, et Holmes juge le risque que A soit vrai suffisamment faible pour repartir sans passer par la cave.

					P=1	P=0				P=1		P=0			
										A=1	A=0	A=1	A=0		
P=1	0.93	A=1	**0.16**	W=1	0.93	0.07	1	H=1	0.09	0.84	0.07	0	1
P=0	0.07	A=0	0.84	W=0	0	0	0	H=0	0	0	0	0	0
P(P)		P(A)		P(W,P)			P(W)	P(H,P,A)					P(H)

TAB. 13 – *Modification du tableau 12 en prenant en compte la certitude que W est vrai*

Nature du réseau bayésien. Le réseau bayésien est une écriture probabiliste de la connaissance d'un expert à un moment donné. Elle est formée de deux parties : la structure du réseau, constituée de règles qui sont des relations de « causes à effets », et codée par des probabilités conditionnelles, et les valeurs de probabilités des évènements (ou faits, propriétés, ou variables). La structure n'est pas susceptible de modifications alors que les probabilités des évènements sont remises à jour dès que l'une d'elles est modifiée pour intégrer une nouvelle connaissance. Les règles du réseau bayésien représentent des relations de type « cause à effet », comme les règles d'implication statistique et les règles logiques vues précédemment, mais leur mise en oeuvre dans un raisonnement est plus complexe : la propagation de l'information d'un nœud à un autre n'est pas simple, elle se fait à l'aide de

[15] D'après cette formule, si X et Y sont deux évènements, on peut écrire P(X|Y)=P(Y|X)P(X)/P(X,Y). Pour le détail de la formule concernant les trois évènements et des calculs associés, nous renvoyons le lecteur intéressé à l'ouvrage de Jensen (1996).

[16] En fait il s'agit ici de P(A=1|W=1,H=1) et non P(A=1) comme indiqué dans le tableau, de la même façon, il faudrait écrire P(A=1|H=1) au lieu de P(A=1) dans le tableau 12.

calculs de mise à jour des probabilités des nœuds qui sont à peine réalisables « à la main », comme nous venons de le voir dans cet exemple très simple.

Nature complexe des nœuds et des règles. Le fait d'avoir un seul arc AB entre deux nœuds de ce réseau ne signifie pas que B est vrai si A est vrai, mais qu'il y a une certaine probabilité que B soit vrai quand A l'est, et une autre quand A ne l'est pas. Si bien que la présence de cet arc dans la figure peut très bien indiquer la règle A→nonB, ou nonA→B, ou nonA→nonB plutôt que A→B, mais pas la règle nonB→nonA qui serait associée à une autre structure contenant à sa place l'arc BA, et avec des probabilités conditionnelles P(A|B) au lieu de P(B|A). Certes, la formule de Bayes permet le passage d'une de ces deux probabilités à l'autre, mais l'une est immuable, fixée avec la structure du réseau, et l'autre change quand les informations permettent de la mettre à jour. Quand ce sont deux arcs qui arrivent à un nœud C, cela indique non pas deux relations correspondant aux règles A→C et B→C mais une relation complexe entre A, B et C, avec dans la majeure partie des cas, une *interaction de A et B sur C*, ce qui signifie que l'action de A sur C diffère selon que B est vrai ou non. C'est le cas dans le tableau 10 du nœud H lié à chacun des nœuds P et A par un arc : la probabilité que la pelouse soit mouillée ne change pas avec le fait que l'arrosage était ou non en marche (1 si A vrai, 1 si A faux) dans le cas ou il a plu (P vrai), alors qu'elle change (0,9 si A vrai, 0 si A faux) quand il n'a pas plu (P faux). Cette relation entre les trois faits est complexe mais elle pourrait l'être plus, car on a supposé pour simplifier les calculs que les deux faits P et A sont indépendants. Dans le cas contraire, en ajoutant un arc de P vers A par exemple, P aurait deux effets sur H, un direct et un indirect par l'intermédiaire de A. Pour limiter la complexité de ces relations, des contraintes sont imposées sous forme de quelques « règles de bonne conduite » selon Naïm et al.(2007)

- se limiter à un nombre raisonnable d'arcs aboutissant à un noeud (pas plus de 4)
- éviter de boucler (même de façon indirecte, un fait ne peut pas être à la fois cause et conséquence d'un autre fait). Il faut contrôler à chaque ajout d'un nouveau lien qu'il ne crée pas de boucle dans le système.
- pas trop de noeuds intermédiaires dans un chemin (pas plus de 4)
- éviter qu'une variable agisse a la fois directement et indirectement sur une autre (bypass, ou dérivation)

Avantages et désavantages de la précision. Nous avons vu que les réseaux bayésiens mettent en lumière des relations complexes, mais précises, alors que l'implication statistique privilégie des relations imprécises, mais simples à appréhender et manipuler. Cette précision en fait un outil privilégié pour l'aide à la décision en économie par exemple. On apprend la structure à partir de données quand on en dispose, et à partir de connaissance expertes sinon, ou même en faisant des hypothèses de travail, puis on le confronte à la réalité en comparant les valeurs obtenues à celles observées, et on corrige si nécessaire. Puis on la manipule en changeant certaines valeurs (par exemple on diminue un taux de prêt, on augmente un investissement, etc.) pour envisager de multiples scénarios. On trouve des études de cas montrant ce fonctionnement dans (Naïm et al. 2007). Il est toutefois un domaine où ce mode de fonctionnement peut être dérangeant, c'est celui de la justice. Un exercice simple proposé dans (Naïm et al. 2007) permet de poser le problème :

> « Un suspect est soupçonné de meurtre. Un témoin fiable à 70% assure l'avoir reconnu, et un test ADN fiable à 99% le désigne comme coupable. Trouver la probabilité a

posteriori qu'il soit coupable sachant que la probabilité a priori qu'il le soit est de 10%. »

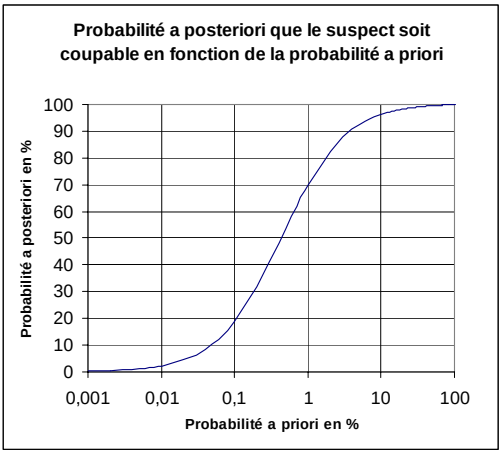

FIG. 7 – *Dépendance entre probabilités a posteriori et celles a priori dans un cas d'école*

On voit dans la figure 7 que la probabilité a posteriori dépend de la probabilité a priori. Les connaissances telles que témoignage et test ADN ne font que la renforcer, par des jeux de multiplication. Si elle est a priori de 10%, elle devient a posteriori à plus de 95%. Les auteurs signalent le problème éthique posé par la fixation de la probabilité a priori.

Au-delà de ce cas d'école, cela montre les limites d'un raisonnement précis sur des données imprécises, aléatoires ou subjectives comme ci-dessus.

Construction de la structure des données Asia. La construction de la structure peut se faire à l'aide d'experts, qui choisissent les nœuds, les arcs, et les valeurs des probabilités a priori et conditionnelles. Cette opportunité est proposée dans les logiciels de réseaux bayésiens. Le graphe de la figure 8, présent dans le logiciel BayesaLab[17] (Version d'évaluation 2001-2008), est un exemple de réseau généré de cette façon.

Cette construction peut aussi se faire par apprentissage à partir des données. Le choix des éléments du réseau se complique dès que la structure augmente de taille. Le plus délicat est de choisir des valeurs précises de probabilités là où on se contenterait d'une échelle ordinale plus ou moins vague (« un peu », « moyennement », « beaucoup ») pour qualifier l'influence de telle variable sur telle autre. La possibilité est offerte dans la plupart de ces logiciels d'apprendre la structure d'après les données. Il y a plusieurs façons de déterminer la structure, la plus « facile », du point de vue algorithmique, étant de faire préciser la structure à l'expert, puis de faire calculer les probabilités selon cette structure par le logiciel d'après les données, et la plus complexe consiste à ne fournir au logiciel que le tableau de données avec les intitulés des variables, qui deviennent toutes des nœuds et laisser le logiciel choisir les arcs et ensuite calculer les probabilités. Pour permettre une solution rapide, des choix par défaut sont faits en suivant des « règles de bonne conduite » comme celles que nous avons

[17] Il est fourni construit dans le logiciel, et un tutoriel explique de façon détaillée comment le construire soi-même (http://www.bayesia.com/fr/produits/bayesialab/).

vues précédemment, mais un paramétrage de ces choix reste possible, détaillé notamment dans BayesaLab par un tutoriel très pédagogique. Nous avons procédé selon la méthode par défaut[18] et obtenu le réseau présent en figure 9.

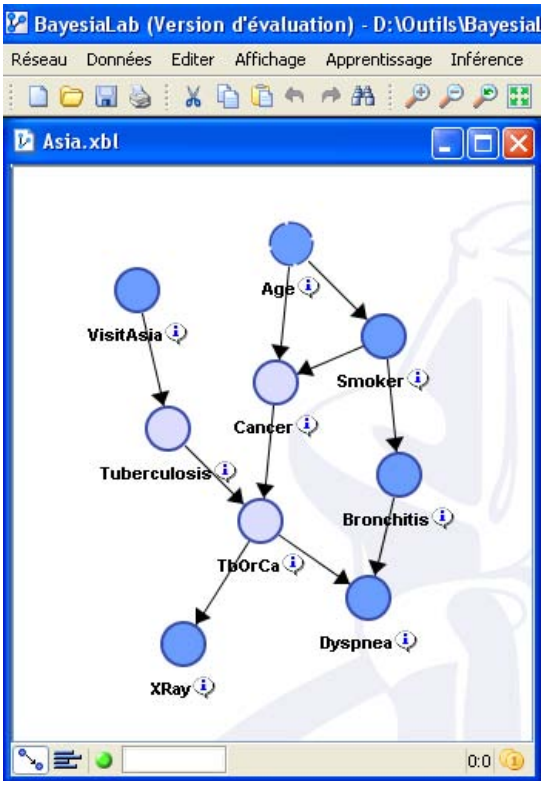

FIG. 8 – *Réseau bayésien construit sur des connaissances expertes.*

Le graphe obtenu par apprentissage total de la structure (cf. figure 9) est très proche de celui proposé par l'expert. Les nœuds sont joints par autant d'arêtes dans les deux figures, seule l'orientation de quelques-unes change : il s'agit des arcs VA→TU, AG→SM, AG→CA qui ont été retournés. Le tutoriel nous invite à intervenir dans le graphe pour le modifier éventuellement en fonction de connaissances expertes, en précisant que l'algorithme n'a pas la possibilité de déceler le sens exact de tous les arcs. Nous avons conservé sans changement le graphe de la figure 9, notre but étant de comparer les « standards » des trois méthodologies. Dans ce graphe, trois nœuds, Age, TbOrCa et

[18] Un choix devait être fait toutefois entre plusieurs méthodes de découverte d'associations (Arbre de recouvrement maximum, Taboo, EQ, SopEQ, Taboo Order). La figure 6 représente le graphe obtenu pour la plupart des options, les autres graphes ayant quelques arcs retournés ou supprimés. Aucun des graphes n'avait les arcs VA→TU, AG→SM, AG→CA dans ce sens, tous les avaient en sens contraire.

Dyspnea, sont plus complexes, étant chacun la destination de deux arcs. De façon plus ou moins importante, les variables Smoker et Cancer interagissent sur la variable Age, Bronchitis et TbOrCa sur la variable Dyspnea, et Cancer et Tuberculosis sur la variable TbOrCa (TbOrCa est l'acronyme de « Tuberculosis Or Cancer », cette variable synthétique a été construite par disjonction des deux autres pour faciliter l'écriture du réseau bayésien).

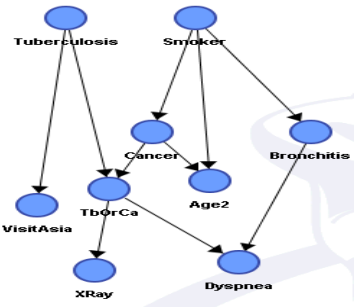

FIG. 9 – *Réseau bayésien construit sur les données Asia (N=7033).*

Comparaison du jeu de règles extrait des données Asia par réseau bayésien aux deux autres. On retrouve dans ce jeu de 10 règles trois règles d'implication statistique de la figure 2, dans le même sens (R1 : Cancer→TbOrCa, R2 Tuberculosis→TbOrCa, R4 TbOrCa→Xray). On peut retrouver les trois règles de ce jeu Smoker→Cancer, Cancer→Age, TbOrCa → Dyspnea dans le graphe d'implication statistique (figure 2, à droite) en s'aidant de la transitivité sous-jacente à ce graphe, la première de ces trois règles étant inversée. Ainsi on retrouve 6 des 10 règles. La règle Smoker→Bronchitis était encore présente dans le tableau 8 de l'indice d'implication non corrigé, et a disparu lors de la correction. L'implication statistique Tuberculosis →VisitAsia ne peut exister, la variable Asia ne donnant pas des relations statistiquement généralisables. Quant à la règle Smoker→Age, qui fait partie de la liaison complexe à trois variables entre Cancer, Smoker et Age du réseau bayésien (cf. figure 9) nous ne la trouvons dans aucun des tableaux de l'implication statistique, alors que nous trouvons à sa place dans le tableau 7 l'implication statistique Smoker→nonAge (valeur 1 de l'indice non corrigé, 0,942 de l'indice corrigé). Cela nous ramène à la remarque déjà faite qu'un arc AB du réseau bayésien indique une relation entre deux variables A et B qui ne peut se réduire à l'écriture de la règle A→B. Par exemple ici, l'arc (Smoker, Age) ne correspond pas à la règle Smoker→Age mais à la règle Smoker→nonAge, alors que les arcs (Smoker, Cancer) et (Cancer, Age) se traduisent par la bonne règle. On voit que le réseau bayésien de la figure 9, apparemment simple, ne peut s'appréhender par une simple lecture du graphique. Ce sont seulement des dépendances qui sont indiquées, et il est nécessaire de consulter les informations chiffrées, ou d'agir sur celles-ci, pour voir si les liaisons correspondantes sont positives ou négatives.

Examinons la règle Tuberculosis →VisitAsia. Sur les 83 personnes atteintes de Tuberculose, et les 74 personnes ayant visité l'Asie, seules 6 vérifient les deux variables. Ce nombre, trop petit d'un point de vue statistique pour générer une règle d'implication statistique ne l'est pas d'un point de vue probabiliste. Dans le premier cas, les sujets sont des unités statistiques, qui, pris individuellement en si petit nombre, ne peuvent être représentatifs d'une tendance, vu la variabilité naturelle des individus. Dans le second cas, 6

personnes sur 74, c'est une part non négligeable d'un sujet abstrait, pouvant être reproduit 74 fois, 7400 fois ou même une infinité de fois. Selon le premier point de vue la liaison entre Tuberculosis et VisitAsia est sans intérêt, alors qu'elle est intéressante selon le second.

Le réseau bayésien est encore plus éloigné du treillis de Galois que l'implication statistique. Bien que le nombre de sujets qui vérifient une règle logique, issue du treillis de Galois puisse être très petit, voire nul, cela ne suffit pas à assurer la ressemblance. Les règles logiques correspondent à des probabilités égales à 1, ce qui ne couvre qu'une très petite partie du réseau bayésien. On ne retrouve que les deux premières des dix règles données par le réseau, celles qui étaient déjà communes au treillis de Galois et à l'implication statistique. Les 15 autres règles logiques ne se retrouvent pas dans le réseau bayésien. La cause la plus vraisemblable de leur absence est le choix fait dans les réseaux bayésiens de ne garder que quelques dépendances très localisées parmi un grand nombre d'indépendances conditionnelles supposées.

6 Conclusion

Nous avons comparé le graphe des implications statistiques selon le modèle de Régis Gras au treillis de Galois et au réseau bayésien. Ce sont trois points de vue différents, statistique, algébrique et probabiliste, pour un même but : donner à l'utilisateur la possibilité de raisonner sur ses données. La brique de base pour chacun est le lien de type causal entre variables, mais la nature des liens et la façon de les enchaîner diffèrent. Les comparaisons ont été faites dans deux directions : aptitude formelle de la représentation à faciliter le raisonnement courant d'un utilisateur et type d'information extraite du jeu de données Asia.

La comparaison des trois ensembles de liens extraits des données Asia montre qu'aucun des modèles ne peut établir automatiquement le sens causal du lien trouvé entre deux faits. Si on sait que A est la cause de B, on peut obtenir la règle A→B comme la règle B→A. La sémantique des liens n'est donc pas directement celle attendue, mais l'utilisateur averti peut, dans le cas du graphe d'implications statistique, interpréter la contraposée nonA→nonB en lieu et place de la règle B→A, et dans le cas des réseaux bayésiens indiquer lui-même le sens attendu.

C'est le modèle à base d'implication statistique qui fournit les liens les plus approximatifs, son but étant de permettre des raisonnements prenant en compte la variabilité individuelle des sujets, comme il est de tradition en statistique. Il offre la navigation la plus simple, des choix drastiques ayant été faits pour que la transitivité manquante soit rétablie, et pour éviter les cycles aboutissant à des « cercles vicieux » du genre A→B, B→C et C→A, qui seraient difficiles à interpréter en terme de causalité. On peut le faire fonctionner sans problème dans les cas limites (variable toujours vraie ou toujours fausse) et ajouter les négations de variables sans obtenir de « fautes de bon sens » du genre A→nonA par exemple. Mais bien sûr, le choix qui a été fait de simplifier sans demander de paramétrage compliqué à l'utilisateur fait perdre certains liens complexes. Des ajouts ont été faits dans les dernières versions, par exemple la possibilité d'avoir plus d'une variable en partie gauche de la règle. Cela permet de faire apparaître des relations complexes entre variables, comme l'interaction (Cadot 2005), mais la représentation qui en est faite s'appuie sur un nombre

parfois important de nœuds contenant plus d'une variable[19], ce qui la destine à un usage expert plutôt que pour le raisonnement courant.

Les treillis de Galois offrent des liens dichotomiques entre deux groupes de variables, qui peuvent être vrais sans même qu'un sujet ne les vérifie, du moment qu'aucun ne les contredit, ce qui en fait un modèle plus adapté aux sciences exactes qu'aux sciences humaines. La rigueur de l'algèbre fait que ce modèle possède toutes les propriétés logiques attendues pour un raisonnement automatique, comme la transitivité par exemple, et le fait de pouvoir engendrer toutes les règles à partir d'une partie génératrice. Il est largement utilisé pour générer les règles d'association (Agrawal 1994), souvent accompagné d'un pré-traitement et d'un post-traitement pour limiter l'explosion combinatoire des règles en éliminant celles qui sont vérifiées par peu de sujets, ou tous « les cas limites » (si les variables x, y et z sont vérifiées par les mêmes sujets, cela fournit un grand nombre de règles sans intérêt) mais en acceptant les règles approximatives. Ces traitements variés ont pour inconvénients de changer sa structure, et de diminuer sa rigueur qui doit être alors renforcée en rajoutant des contraintes qui le rendent difficile à paramétrer pour l'adapter aux besoins de l'utilisateur final. A notre connaissance, aucune amélioration n'a été faite pour rendre ce modèle compatible avec la négation des variables.

Les réseaux bayésiens fournissent des liens très précis, soumis à un certain nombre de contraintes, appelées « règles de bonnes conduite », afin de rendre le réseau plus simple, mais la navigation ne peut se faire au simple vu du réseau, même s'il ne comporte que peu de variables. Les liens sont en effet complexes et l'information circule ou est bloquée selon que l'information dans les nœuds est ou non connue et suivant la façon dont les liens se succèdent. Cette modélisation ne fournit pas de connaissance abstraite sur les données, simple à manipuler pour un cerveau humain, mais de la connaissance concrète, bien appréciée dans les sciences économiques et de gestion pour simuler des faits comme des choix d'investissements, et prendre des décisions optimales pouvant être justifiées de façon comptable. A noter que dès qu'on s'écarte un peu de ce domaine très quantitatif et qu'on se rapproche des sciences humaines hors du domaine économique, on ne peut que constater que cette mécanique de haute précision est mal adaptée aux faits humains. L'exemple qui est fourni dans le domaine de la justice montre comment la probabilité qu'un accusé soit coupable dépend de probabilités variées qui sont établies a priori, c'est-à-dire provenant d'un expert, ou qui sont estimées d'après des données collectées. Ces deux sources sont susceptibles d'erreurs, dues à la subjectivité d'un individu ou à la variabilité d'un groupe d'individus, qui ne sont, à notre connaissance, pas prises en compte par les modèles de réseaux bayésiens.

Références

Agrawal R. and R. Srikant (1994), *Fast algorithms for mining association rules in large databases*, Research Report RJ 9839, IBM Almaden Research Center, San Jose, California.

[19] Sur les données « Asia », l'ajout de la possibilité d'avoir 2 variables en partie gauche des règles a presque quadruplé le nombre de nœuds et le nombre de règles.

Graphe de règles d'implication statistique pour le raisonnement courant.

Barbut M. et B. Monjardet (1970), *Ordre et classification*, Tome 2, Paris, Hachette.

BayesiaLab - Evaluation version (c) Copyright Bayesia S.A. 2001-2008 http://www.bayesia.com/fr/produits/bayesialab/release/bayesialab-3-3.php

Ben Naceur-Mourali, C. Gonzales (2004). Une unification des algorithmes d'inférence de Pearl et de Jensen. *Revue d'intelligence artificielle. RSTI série RIA*, Vol 18, no 2/2004. Lavoisier, Paris. 229-260

Birkhoff, G. (1948). *Lattice theory*, American Mathematical Society colloquium publications volume 25. New York.

Cadot, M. (2006). *Extraire et valider les relations complexes en sciences humaines : statistiques, motifs et règles d'association*. Thèse de doctorat en informatique. Université de Franche-Comté.

Cadot, M., Maj, J.-B. and Ziadé T. (2005), Association Rules and Statistics, in *Encyclopedia of Data Warehousing and Mining*, édité par John Wang, Montclair State University, USA, 94-98.

Davey B.A., Priestley H.A. (1990) *Introduction to Lattices and Order*, Cambridge University Press.

Formal Concept Analysis (2009): http://www.upriss.org.uk/fca/fca.html

Godin R., Mineau G., Missaoui R., Mili H. (1995). Méthodes de classification conceptuelle basées sur les treillis de Galois et applications, *Revue d'Intelligence Artificielle*, 9(2), 105-137

Gras R. (1979). *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs, didactiques en mathématiques*, Thèse de doctorat, Université de Rennes I.

Gras R., Bailleul M. (2000). La fouille dans les données par la méthode d'analyse implicative statistique, *Journées du 23 et 24 juin 2000 organisées par l'IUFM de Caen et l'ARDM*.

Gras R., P. Kuntz, R. Couturier, F. Guillet (2001). Une version entropique de l'intensité d'implication pour les corpus volumineux, *Extraction des connaissances et apprentissage*, 2001, Volume 1, Numéro 1-2, 69-80.

Guigues J.L. et V. Duquenne (1986). Familles minimales d'implications informatives résultant d'un tableau de données binaires, *Math. Sci. Hum*. n°95, 5-18

Guillet F. (2004). Mesure de qualité des connaissances en ECD, *Tutoriel EGC 2004*, Clermont-Ferrand, 20 janvier 2004.

Jensen F.V. (1995). *An introduction to Bayesian Networks*. University College London.

Mephu Nguifo E**.** (1994). Une nouvelle approche basée sur le treillis de Galois pour l'apprentissage de concepts, *Mathématiques, Informatique, Sciences Humaines*, vol. 134, 19-38,

Morineau, A. (éd.) (1995). *Aide-mémoire statistique.*. Saint-Mandé : CISIA-CERESTA

Naïm P., P. H. Wuillemin, P. Leray, O. Pourret et A. Becker (2007). *Réseaux bayésiens*, Collection Algorithmes, Eyrolles, Paris.

Muller P ; et L. Vieu. (2009) Structures d'ordre, dans *Sémanticlopédie, dictionnaire de sémantique*, http://www.semantique-gdr.net/dico/index.php/Structures_d'ordre

Pearl J. (2000) *Causality models, reasoning, and inference*, Cambridge University Press, 267 - 279.

Reinhardt F. et H. Soeder (1974) (édition française de J. Cuenat de 1997). Atlas des mathématiques. Collection la pochothèque. Librairie Générale Française..

Whittaker J. (1990). *Graphical models in applied multivariate Statistics*, John Wiley.

Annexes

Un treillis peut être défini de deux façons, soit par une relation d'ordre (cf. définition 1), soit pas deux lois internes (cf. définition 2). Et ces deux définitions sont équivalentes (Reinhardt et al. 1974).

Définition 1 : Un treillis (E, R) est un ensemble E muni d'une relation d'ordre R, c'est-à-dire une relation binaire sur E *réflexive*, *transitive* et *antisymétrique*, telle que toute paire d'éléments admette une *borne supérieure* et une *borne inférieure*. Voici les définitions de ces termes, dans lesquelles b, x, y et z sont des éléments de E :

- réflexivité : $\forall x, (x\ R\ x)$
- antisymétrie : $\forall x, y,$ si $(x\ R\ y)$ et $(y\ R\ x)$ alors $(x = y)$
- transitivité : $\forall x, y, z,$ si $(x\ R\ y)$ et $(y\ R\ z)$ alors $(x\ R\ z)$
- b est une borne inférieure de {x, y}si :
  - $(b\ R\ x)$, $(b\ R\ y)$,
  - $\forall z,$ si $(z\ R\ x)$ et $(z\ R\ y)$ alors $(z\ R\ b)$

La relation réciproque R' de R (x R'y est définie par y R x) est également une relation d'ordre sur E. Ce qui permet de définir la borne supérieure pour R de deux éléments comme étant la borne inférieure pour R'.

Un treillis est *complet* si tout sous-ensemble de E admet un plus petit élément et un plus grand élément. C'est le cas notamment des treillis construits sur des ensembles finis.

Définition 2 : Un treillis $(E, \wedge, \vee)$ est un ensemble muni de deux lois internes possédant un certain nombre de propriétés : *commutativité*, *associativité*, *absorption*. Voici les définitions de ces termes, dans lesquelles x, y et z sont des éléments de E :

- commutativité : $\forall x, y,\ (x \vee y) = (y \vee x)$; $(x \wedge y) = (y \wedge x)$
- associativité : $\forall x, y, z, ((x \vee y) \vee z) = (x \vee (y \vee z))$; $((x \wedge y) \wedge z) = (x \wedge (y \wedge z))$
- absorption : $\forall x, y, (x \wedge (x \vee y)) = y$; $(x \vee (x \wedge y)) = y$

Un treillis *booléen* est un treillis possédant des propriétés supplémentaires : il a un *élément nul*, un *élément universel*, ses lois sont *distributives* l'une par rapport à l'autre, et ont la propriété de *complémentation*. Voici les définitions de ces termes, dans lesquelles x, y et z sont des éléments de E :

Graphe de règles d'implication statistique pour le raisonnement courant.

- distributivité de la loi $\vee$ par rapport à la loi $\wedge$: $\forall x, y, z, (x \vee (y \wedge z)) = ((x \wedge y) \vee (x \wedge z))$
- distributivité de la loi $\wedge$ par rapport à la loi $\vee$: $\forall x, y, z, (x \wedge (y \vee z)) = ((x \vee y) \wedge (x \vee z))$
- n élément nul : $\forall x, (n \wedge x) = n$; $(n \vee x) = x$
- u élément universel $\forall x, (u \wedge x) = x$; $u \vee x) = u$
- complémentation : $\forall x, \exists x', (x \wedge x') = n, (x \vee x') = u$

Un tel treillis est aussi appelé treillis distributif complémenté.

Equivalence des deux définitions : on passe du treillis (E, R) au treillis (E, $\wedge$, $\vee$) en définissant pour loi $\wedge$ (respectivement $\vee$) l'application de ExE vers E qui à toute paire d'éléments {x,y} de E associe leur borne inférieure (resp. supérieure). Et on passe dans le sens inverse en définissant la relation R pour toute paire d'éléments {x,y} de E par (x R y) si x est la borne inférieure de {x,y} (ou, ce qui revient au même, si y est la borne supérieure de {x,y}). Le lecteur intéressé par la démonstration de cette équivalence peut consulter les ouvrages relatifs aux treillis (Birkhoff 1948). On note (E, $\wedge$, $\vee$, R) le treillis pour lequel les lois et la relation d'ordre sont ainsi associées.

Treillis d'ensemble. Un ensemble E étant donné, le *treillis d'ensemble* est une structure de treillis sur l'ensemble de toutes les parties de E (on peut le noter 2^P), les lois internes étant l'intersection et la réunion ensemblistes, et la relation d'ordre étant l'inclusion ensembliste. C'est un treillis booléen dont l'élément nul est l'ensemble vide et l'élément universel l'ensemble E. On peut établir aisément que ce sont respectivement le plus petit élément et le plus grand élément du treillis. D'après P. Muller et L. Vieu. (2009), on obtient une structure d'algèbre de Boole, qui est en correspondance directe avec la structure des règles logiques, « Boole a montré que l'ensemble des parties d'un ensemble muni de l'inclusion est isomorphe au calcul des propositions en logique : l'élément universel E correspond à Vrai, $\varnothing$ à Faux, l'inclusion à l'implication, l'intersection à la conjonction, l'union à la disjonction et l'opérateur de complément à la négation ».

Summary

The statistical implication rules may look similar to the rules of mathematical reasoning. However the model underlying the former is not the formal logic model underlying the latter; it is a statistical model giving rise to approximate relations, in accordance with the common sense logic. The problem is that chaining approximate rules may yield inconsistencies, unless justifiable rule connections may be available in a graph. We will show in this chapter how to set up this chaining in order to keep clear of this drawback, particularly through the building of the implicative graph such as proposed by the successive versions of CHIC. We will compare this statistical model of the data to two alternative models: one is an algebraic model, the Galois lattice, the other one is probabilistic, i.e. bayesian networks. For the ease of comparison, we will illustrate the operation of the three models through a medical dataset freely accessible via internet.