Apprentissage de la structure des réseaux bayésiens à partir des motifs fréquents corrélés : application à l'identification des facteurs environnementaux du cancer du Nasopharynx

Alexandre Aussem*, Zahra Kebaili*, Marilys Corbex**, Fabien De Marchi***

*Equipe COMAD, Lab. PRISMa, Université Lyon 1,
alexandre.aussem@univ-lyon1.fr,
**Unité d'épidémiologie génétique,
Centre International de Recherche sur le Cancer (CIRC), Lyon,
corbex@iarc.fr,
***LIRIS UMR CNRS 5205, Université Lyon 1,
fabien.demarchi@liris.cnrs.fr

Résumé. L'apprentissage de structure des réseaux bayésien à partir de données est un problème NP-difficile pour lequel de nombreuses heuristiques ont été proposées. Dans cet article, nous proposons une nouvelle méthode inspirée des travaux sur la recherche de motifs fréquents corrélés pour identifier les causalités entre les variables. L'algorithme opère en quatre temps : (1) la découverte par niveau des motifs fréquents corrélés minimaux ; (2) la construction d'un graphe non orienté à partir de ces motifs ; (3) la détection des V_structures et l'orientation partielle du graphe ; (4) l'élimination des arêtes superflues par des tests d'indépendance conditionnelle. La méthode, appliquée au réseau *Asia*, permet de retrouver la structure du graphe initial. Nous l'appliquons ensuite aux données d'une étude épidémiologique cas-témoins du cancer du nasopharynx (NPC). L'objectif est de dresser un profil statistique type de la population étudiée et d'apporter un éclairage utile sur les différents facteurs impliqués dans le NPC.

1 Introduction

Les réseaux d'inférence bayésiens (RB) sont des outils d'apprentissage numérique qui permettent de rendre compte de relations causales entre des variables aléatoires et de construire un raisonnement probabiliste à partir de connaissances, parfois incertaines et incomplètes, consignées dans les bases de données. L'apprentissage automatique des *valeurs numériques* des probabilités conditionnelles s'opère d'ordinaire à partir d'un ensemble d'apprentissage, même incomplet, si la structure du réseau est *connue*. En revanche, l'apprentissage de la *structure* du RB à partir de données est plus problématique ; la taille de l'espace de recherche est super-exponentielle en fonction du nombre de variables et le problème combinatoire associé est NP-difficile. Deux grandes familles de méthodes existent : celles fondées sur la recherche de causalités via des tests d'indépendance conditionnelle et celles fondées sur la maximisation d'un score. Avec les méthodes à base de score, l'ajout d'un arc repose sur un compromis entre

augmentation de la vraisemblance et pénalité associée à cet arc. En conséquence, les relations causales dans lesquelles la cause apparaît rarement sont souvent ignorées. En outre, l'inférence d'un modèle causal *complet* est réputée impossible (Silverstein et al., 1998) lorsque le nombre de caractère devient prohibitif (de l'ordre de plusieurs milliers) car ces heuristiques sont facilement piégées dans les nombreux minima locaux.

Dans cet article, l'accent est mis sur la découverte de causalités par des méthodes basées sur des contraintes (Cooper, 1997; Friedman et al., 1998; Pearl, 2000; Spirtes et al., 2000). Les contraintes imposées à la structure du graphe causal proviennent des informations statistiques sur les dépendances et indépendances conditionnelles observées. Deux séries d'algorithmes ont été proposées, l'algorithme IC (*Inductive Causation*) (Pearl et Verma, 1991; Pearl, 2000) et les algorithmes SGS, PC (Spirtes et al., 2000). Elles reposent sur l'hypothèse que toutes les variables d'intérêt sont connues (*suffisance causale*), c'est aussi l'hypothèse que nous ferons dans cet article. Ces algorithmes commencent par construire un graphe non orienté en rajoutant (ou supprimant) au fur et à mesure les arêtes à l'aide des tests d'indépendance conditionnelle, cherchent les V-structures puis propagent l'orientation sur les arêtes adjacentes. La recherche d'indépendances conditionnelles étant l'étape la plus coûteuse. Pour réduire le nombre a priori exponentiel de tests d'indépendance en deux variables A et B conditionnellement à un ensemble de variables S, des heuristiques ont été proposées ; PC procède par niveau en considérant les indépendance conditionnelle d'ordre 0, 1, 2 etc. Ces heuristiques reposent sur un choix judicieux des variables dans S (e.g. restrictions aux variables adjacentes à A et B, parcours dans l'ordre lexicographique ou selon l'ordre de la statistique choisie pour le test d'indépendance). Ce parcours par niveau, et l'ordre arbitraire des variables conditionnelles, n'est pas sans rappeler les problématiques rencontrées dans la recherche des motifs fréquents.

Cet article présente une nouvelle heuristique fondée sur la recherche de causalités en exploitant des algorithmes de recherche de motifs. Les algorithmes d'extraction de motifs et de règles ont fait l'objet de nombreux travaux de recherche ces 10 dernières années ; ce problème constitue encore l'une des applications les plus populaires de l'extraction de connaissances dans les données, ou fouille de donnée (Han et Kamber, 2000; Hand et al., 2001). L'objectif principal annoncé dans cette thématique est le *passage à l'échelle*, c'est-à-dire l'implantation d'algorithmes capables de traiter des jeux de données réels en temps raisonnable, d'où leur intérêt pour identifier la structure des RB. La méthode proposée dans cet article s'inspire de ces travaux, et plus particulièrement des travaux sur la recherche de motifs fréquents corrélés par des algorithmes par niveau. L'objectif est *in fine* d'identifier les causalités entre des groupes de variables. Les algorithmes d'extraction des motifs reposent généralement sur un parcours efficace de treillis et sont facilement parallélisables (Adamo, 2001). Ils bénéficient en outre d'efforts de recherche particuliers concernant les structures de données, la gestion de la mémoire, ainsi que l'accès aux bases de données. C'est donc dans l'optique d'améliorer l'efficacité des heuristiques que nous proposons de marier les techniques de fouilles à l'apprentissage de la structure du RB. L'algorithme opère selon le même schéma que IC : découverte par niveau des motifs fréquents corrélés minimaux ; construction d'un graphe non orienté à partir de ces motifs ; détection des V_structures ; orientation partielle du graphe, et élimination des arêtes superflues par des tests d'indépendance conditionnelle.

L'algorithme proposé a été développé sous Matlab à l'aide de la Toolbox BNT de (Murphy, 2001) et de la Toolbox BNT-SLP de Leray et Francois (2004). Appliqué au réseau de la dyspnée *Asia* (Lauritzen et Speigelhalter, 1988), nous montrons qu'il permet de retrouver la

structure du graphe initial à partir de 15000 données. C'est un résultat encourageant comme le montre l'évaluation critique des algorithmes d'apprentissage de structure des RB présenté dans (Francois et Leray, 2004). Après cette rapide validation sur ce cas d'école, nous appliquons ensuite notre heuristique aux données d'une étude épidémiologique cas-témoins du cancer du nasopharynx (NPC). L'objectif est de dresser un profil statistique type de la population étudiée et d'apporter un éclairage utile sur les différents facteurs impliqués dans le NPC.

2 Préliminaires

2.1 Des règles d'association aux motifs corrélés

La découverte de règles d'association est l'une des tâche de fouille de données les plus étudiées depuis son introduction par (Agrawal et al., 1993). (Agrawal et Srikant, 1994) propose une approche en deux temps : la découverte des motifs dits "fréquents" à partir d'un seuil défini par l'utilisateur, puis la construction des règles proprement dites. La première étape est alors réputée la plus difficile, puisqu'elle constitue la phase d'accès aux données ; les plus grands efforts de recherche ont été consacrés à ce sous-problème avec pour objectif le "passage à l'échelle" aux gros volumes de données réels. Des algorithmes efficaces ont été alors proposés pour réduire, en pratique, l'explosion combinatoire du parcours de l'espace de recherche. Le plus populaire, et qui reste encore souvent le seul algorithme implanté dans les outils commerciaux de fouille de données, est l'algorithme $APriori$ (Agrawal et Srikant, 1994). Celui-ci exploite la propriété d'anti-monotonie du prédicat "être fréquent" qui permet, lors d'un parcours "par niveau" de l'espace de recherche, d'élaguer un grand nombre de candidats potentiels. De très nombreux algorithmes on été proposés par la suite, que ce soit en introduisant des propriétés supplémentaires améliorant l'élagage (Pasquier et al., 1999; Bastide et al., 2000), en changeant le mode de parcours (Bayardo, 1998; Gunopulos et al., 2003), ou encore en optimisant les structures de données utilisées (Han et al., 2000).

Au vu de l'efficacité de ces algorithmes, il est tentant de s'inspirer des règles d'association pour construire la structure d'un graphe causal. Toutefois, il s'agit de règles logiques, bien différentes d'une relation probabiliste entre deux variable aléatoires. Les règles portent sur des éléments $\{A = a_j\}$, avec a_j la j-ième modalité du caractère A, tandis que les nœuds d'un RB représentent les caractères eux-mêmes. Dans ces conditions, comment exploiter l'efficacité des algorithmes de découverte des motifs fréquents, sachant que les motifs qui nous intéressent sont désormais des ensembles de variables aléatoires et non plus des ensembles d'événements ? Deux solutions existent. La première est de travailler avec les motifs (d'événements) fréquents pour ultérieurement appliquer des critères de pertinence de règles d'association proches de la notion de causalité probabiliste. Un aperçu assez complet des mesures d'intérêt proposées peut être trouvé dans (Bayardo et Agrawal, 1999). La seconde solution est d'extraire directement des motifs de variables aléatoires corrélées. Il faut pour cela redéfinir la notion de support et définir une mesure de corrélation. Parmi ces mesures possibles, le χ^2 présente des propriétés intéressantes. Il est en effet possible dans le cas de variables booléennes d'exploiter une propriété de monotonie qui, lors d'un parcours "par niveau" de l'espace de recherche, permet de se concentrer sur les ensembles fréquents corrélés *minimaux* dans le treillis des parties. C'est donc cette option que nous prenons.

2.2 Les motifs corrélés minimaux

Il faut dans un premier temps clarifier la notion de corrélation entre caractères. Soit $p(a)$ la probabilité qu'un événement a se produise et $p(\overline{a}) = 1 - p(a)$ celle de l'événement contraire. $p(\overline{a}b)$ est la probabilité que b se produise mais pas a. Les événements a et b sont dits indépendants si $p(ab) = p(a)p(b)$. De même, si $p(abc) = p(a)p(b)p(c)$ alors a,b et c sont indépendants. Si l'un des ab, $\overline{a}b$, $a\overline{b}$, $\overline{ab}$ est dépendant alors les variables aléatoires A et B sont dépendantes. De même si l'une des huit combinaisons de 3 événements $(abc, ab\overline{c}, \ldots, \overline{a}\overline{b}\overline{c})$ sont dépendants alors la variables aléatoires A,B et C sont dépendantes et ainsi de suite pour des ensembles plus grands. Dans la suite, soit I un ensemble de variables aléatoires ; nous appellerons *motif* un sous-ensemble de I. Soit D une base de données constituée d'un ensemble d'expériences, pour lesquelles chaque variable $A \in I$ prend la valeur a ou $\overline{a}$. Une des propriétés importantes de la dépendance est d'être monotone dans l'ensemble des parties de I ordonné par l'inclusion. Si X est un motif dépendant, alors tous ses sur-ensembles sont dépendants. Dans ce contexte, l'ensemble des motifs dépendants peut-être exactement représenté par ses éléments minimaux. Toutefois, la notion de dépendance est théorique, dans la pratique seule une mesure de *corrélation* peut nous renseigner sur le degré de dépendance entre des caractères. La statistique du χ^2, que nous décrivons dans la suite, permettra de mesurer le degré de corrélation d'un motif. Nous montrons dans ce qui suit que cette mesure empirique est également monotone dans l'ensemble des parties de I ordonné par l'inclusion.

2.3 La statistique du χ^2

Le test du χ^2 est couramment utilisé dans les tests d'adéquation de loi et les tests d'indépendance. Dans ce qui suit, nous rappelons brièvement les propriétés de la statistique du χ^2 et son utilisation pour la découverte d'ensembles de variables corrélées minimaux. Considérons deux caractères quelconques A et B et un échantillon de taille n issu d'une population. Si l'on suppose que A et B possèdent respectivement a et b modalités, on note n_{ij} le nombre d'observations appartenant à la i-ème modalité de A et à la j-ème modalité de B, puis $n_{i.} = \sum_{j=1}^{a} n_{ij}$ et $n_{.j} = \sum_{j=1}^{b} n_{ij}$. La statistique du χ^2 permet de tester l'hypothèse H_0 "les deux caractères sont indépendants" contre H_1 "les deux caractères sont dépendants" au risque α. Pour cela on calcule les effectifs théoriques $n'_{ij} = n_{i.}n_{.j}/n$ sous l'hypothèse d'indépendance et l'on applique la formule du χ^2_{AB} ci-dessous. On rejette H_0 si $\chi^2_{AB} > \chi^2_{1-\alpha}(\nu)$ où $\chi^2_{1-\alpha}(\nu)$ est le quantile d'ordre $1-\alpha$ de la loi du χ^2 à $\nu = (a-1)(b-1)$ degrés de liberté. En pratique cette approximation n'est utilisée dès et que $n > 30$, $n_{ij} > 5$ et $n'_{ij} > 5$, dans le cas contraire, il faut procéder à des regroupements de modalités voisines. L'intervalle de rejet $[\chi^2_{1-\alpha}(\nu), +\infty[$ est certes arbitraire (les faibles écarts $\chi^2_{AB} << \chi^2_{1-\alpha}(\nu)$ sont également rares) mais il capture la notion intuitive de dépendance. Dans le cas de 3 variables A, B et C, χ^2_{ABC} s'obtient à partir des $n'_{ijk} = n_{i..}n_{.j.}n_{..k}/n^2$. χ^2_{ABC} suit approximativement une loi χ^2 à $\nu = (a-1)(b-1)(c-1)$. Dans notre étude, le χ^2_S sert de mesure de corrélation entre un ensemble S de caractères. Son avantage par rapport aux mesures classiques de règles d'association est de prendre en compte simultanément toutes les cellules de la table de contingence et définir des seuils ad hoc à partir de tables statistiques. On montre facilement dans le cas binaire que la propriété de fermeture reste valable pour la statistique du χ^2 : si S est corrélé au risque α, tout sur-ensemble est encore corrélé au risque α (e.g. $\chi^2_{ABC} > \chi^2_{AB}$). De plus, $\nu = 1$ dans le cas binaire, ce qui permet de comparer entre elles les valeurs du χ^2_{ABC} et de fixer un unique

seuil de confiance à un risque donné. Nous serons également amenés à mesurer le degré de dépendance *conditionnelle* entre deux variables, $\chi^2_{AB|C}$ avec $n'^k_{ij} = n_{i.k}n_{.jk}/n_{..k}$. $\chi^2_{AB|C}$ suit une loi χ^2 à $\nu = (a-1)(b-1)c$ degrés de liberté. On observe, dans le cas binaire, que le nombre de degrés de liberté de $\chi^2_{AB|S}$ où S est un ensemble de n variables binaires varie cette fois en 2^n. χ^2_{AB}, χ^2_{ABC} et $\chi^2_{AB|C}$ sont donnés par les formules :

$$\chi^2_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{(n_{ij}-n'_{ij})^2}{n'_{ij}},\ \chi^2_{ABC} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\frac{(n_{ijk}-n'_{ijk})^2}{n'_{ijk}},\ \chi^2_{AB|C} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\frac{(nijk-n'^k_{ij})^2}{n'_{ijk}}$$

2.4 Dépendance, indépendance et degré de corrélation

Dans le suite, nous jonglerons avec plusieurs seuils, $\chi^2_{dep}(\nu)$ et $\chi^2_{indep}(\nu)$, associés à la loi du χ^2 à ν degrés de liberté, pour tester la dépendance ou l'indépendance entre des ensembles de variables. Ainsi pour tout ensemble S de variables, $\chi^2_S > \chi^2_{dep}(1)$ indiquera la dépendance et $\chi^2_S < \chi^2_{indep}(1)$ indiquera l'indépendance, avec $\chi^2_{indep}(1) \leq \chi^2_{dep}(1)$. Dans le cas où $\chi^2_{indep}(1) < \chi^2_S < \chi^2_{dep}(1)$, l'ensemble S sera donc jugé ni dépendant, ni indépendant, selon le test qui lui est apliqué. De même, $\chi^2_{AB|S} < \chi^2_{indep}(2^{card(S)})$ indiquera l'indépendance conditionnelle de A et B sachant un ensemble S de variables. On prendra pour $\chi^2_{dep}(\nu)$ le quantile d'ordre $1-\alpha$ (avec $\alpha = 0.05$) de la loi du χ^2 à ν degrés de liberté ; l'autre seuil $\chi^2_{indep}(\nu)$ est arbitraire car il n'est pas possible d'estimer la probabilité de mal classer des variables dépendantes (voir (Silverstein et al., 2000) pour une discussion sur les seuils). Pour comparer le degré de corrélation entre AB et $AB|C$, la comparaison des valeurs du χ^2 n'est plus possible puisque χ^2_{AB} et $\chi^2_{AB|S}$ suivent toutes deux une loi χ^2 mais avec des degrés de liberté distincts. Pour contourner ce problème, nous proposons le test suivant : $AB|C$ seront jugés plus corrélés que AB dès lors que $F_{\chi^2(2)}(\chi^2_{AB|C}) > F_{\chi^2(1)}(\chi^2_{AB})$ où $F_{\chi^2(\nu)}$ est la fonction de répartition d'un χ^2 à ν degrés de liberté (i.e. $F_{\chi^2(\nu)}(x) = P(\chi^2(\nu) \in]-\infty, x])$). Indépendance et corrélation sont deux notions intimement liées, car dans le cas de deux variables booléennes A et B, on montre aisément que $\chi^2_{AB} = n\rho^2$ où ρ est le coefficient de corrélation entre A et B, et n la taille des données (Silverstein et al., 2000). La force de la corrélation est donc directement mesurée par la statistique du χ^2. C'est pourquoi l'estimation du degré de dépendance mesuré par le χ^2 revient ici à estimer le degré de corrélation entre les variables booléennes.

3 Une nouvelle heuristique pour l'apprentissage de la structure du RB

Notre méthode d'apprentissage de structure de RB opère en plusieurs étapes. La première partie consiste en la découverte de corrélations entre variables à partir de données. La seconde partie traite de la construction du graphe non orienté qui est sensé, une fois les arêtes dirigées, représenter la structure du réseau bayésien résultant de l'apprentissage. La dernière partie consiste à diriger les arêtes du graphe non orienté et supprimer celles qui sont superflues. Pour illustrer et valider notre approche, nous avons utilisé le réseau de la dyspnée *Asia* (Lauritzen et Speigelhalter, 1988), dont la structure est déjà connue (voir Figure 1).

3.1 Découverte des motifs corrélés minimaux fréquents

Nous décrivons ici succinctement la découverte de motifs corrélés minimaux fréquents telle qu'elle a été définie et réalisée dans (Silverstein et al., 1998), donc l'objectif était de généraliser des règles d'association en règles de corrélation. Cette phase d'accès aux données est la plus coûteuse en terme de complexité. Le permier algorithme a pour objectif d'extraire l'ensemble des motifs corrélés minimaux, qui sont, par monotonie, une représentation de tous les motifs corrélés. En outre, une notion de fréquence est également utilisée afin d'élaguer au mieux les ensembles inutiles. Toutefois, la notion classique de support dans le domaine de la découverte de règles d'associations ne s'applique qu'aux modalités des caractères et ne caractérise pas une relation probabiliste entre les variables aléatoires. Pour tenir compte de toutes les cellules de la table de contingence, une généralisation de la notion de support s'impose : un motif X sera dit fréquent dès lors qu'au moins $p\%$ des cellules de la table de contingence ont un support supérieure à s, p et s étant défini par l'utilisateur. Notons que cette définition permet de conserver la propriété d'anti-monotonie du support, sur laquelle repose en partie l'efficacité des méthodes d'extraction. En combinant la nouvelle mesure du support avec celle du χ^2, on obtient l'algorithme 1 "$\chi^2 - support$" de (Silverstein et al., 1998) non reproduit ici faute de place. L'algorithme parcourt par niveau l'espace de recherche et retourne la liste SIG des fréquents corrélés minimaux tels que tous les sous-ensembles de taille $i-1$ des éléments de taille i de SIG sont fréquents mais non-corrélés. La complexité au pire cas de l'heuristique proposée est *exponentielle* en fonction du nombre de caractères. En pratique toutefois, l'efficacité dépend de la sélectivité des prédicats d'élagage dont l'étude est encore à ce jour un problème ouvert. Dans nos expérimentations préliminaires, le nombre de niveaux i explorés n'a jamais dépassé deux. Une étude empirique sur des bancs d'essais permettra de comparer cette heuristique avec les autres sur la base du temps de calcul effectif.

3.2 Construction du graphe non orienté

Après avoir extrait les ensembles fréquents corrélés minimaux, se pose la question de la représentation des variables par un graphe causale non orienté. Relier toutes les variables corrélées entre elles n'est pas une solution. Si $A \rightarrow B \rightarrow C$ où "'$\rightarrow$"' désigne une relation causale, alors A et C seront certainement corrélées par transitivité. Il faut donc s'interroger sur l'existence réelle de l'arête (A, C) dans le graphe et éliminer certains couples de caractères corrélés sans causalité directe. Dans ce but, nous appliquons la proposition suivante pour tester si la relation est de type CCC Cooper (1997) :

Proposition 1 *Pour tous* $(A, B), (B, C) \in E$, *si* $\chi^2(AC|B) < \chi^2_{indep}(2)$ *alors* $(A, C) \notin E$

L'algorithme 2 (non représenté faute de place) traite dans un premier temps les motifs corrélés portant sur des couples de variables. Il utilise l'ensemble des couples corrélés ordonnés dans l'ordre décroissant du χ^2. Pour tout couple corrélé (A, B), plusieurs cas peuvent se présenter : (1) s'il n'existe aucun chemin, à ce stade de l'algorithme, entre A et B, dans G alors l'arête (A, B) est ajoutée à E ; (2) s'il existe un chemin de la forme $A - C - B$, alors on applique la proposition 1. Si $\chi^2(AB|C) < \chi^2_{indep}(2)$, l'arête (A, B) n'est pas ajoutée à E ; (3) Dans les autres cas, on ajoute (A, B) à l'ensemble des arêtes dites *potentielles*. Il faut attendre de connaître la nature des chemins après la détection des V-structures pour statuer sur l'existence de l'arc.

Pour les motifs corrélés d'ordre supérieur à deux, disons $ABCD$, nous avons choisi arbitrairement de créer une V-structure avec une variable cible, disons B, telle que $\chi^2(ACD|B)$ soit maximal, et de relier A, B et C à B. L'algorithme 1 et 2 ont été appliqués au réseau *Asia* avec les paramètres suivants : pourcentage du support $p = 25\%$, $\chi^2_{indep}(1) = \chi^2_{dep}(1) = 3.84$ et $\chi^2_{indep}(2) = 5.99$. 15000 exemples ont été échantillonnés grâce au réseau théorique. La Figure 2 montre le graphe non orienté obtenu à l'issue cete étape avec les valeurs précédentes. On observe que toutes les arêtes du réseau original *Asia* ont bien été selectionnées mais que 5 arêtes supplémentaires figurent à tort sur le graphe.

3.3 Détection des V-structures

Comment diriger les arêtes du graphe non-orienté ? Il existe 3 types de connexions : les connexions en série $A \rightarrow B \rightarrow C$ ou $A \leftarrow B \leftarrow C$, la connexion divergente $A \leftarrow B \rightarrow C$ et la connexion convergente $A \rightarrow B \leftarrow C$. Ces dernières sont encore appelées V-structures. Il faut garder à l'esprit que plusieurs structures peuvent encoder la même loi de probabilité conjointe (équivalence de Markov). Seules les V-structures, détectées par les tests d'indépendance conditionnelle, permettent de diriger certains arcs et de propager arbitrairement les orientations des autres arcs. Considérons une V-structure $A \rightarrow B \leftarrow C$, alors, conditionnellement à B, la relation de corrélation entre A et C est renforcée. On s'attend donc à $F_{\chi^2_2}(\chi^2(AC|B)) > F_{\chi^2_1}(\chi^2(AC))$, quelque soient les autres chemins entre A et C non représentés. Nous ne testons pas l'indépendance entre A et C car ils peuvent ne pas être indépendants à cause de parents communs comme le sont par exemple O et B dans le réseau *Asia*. A l'inverse, si $A \leftarrow B \leftarrow C$, $A \rightarrow B \rightarrow C$ ou $A \leftarrow B \rightarrow C$ alors, conditionnellement à B, la corrélation entre A et C ne peut que diminuer, quelque soient les autres chemins entre A et B. D'où la proposition suivante :

Proposition 2 *Soit A, B, C trois variables telles que $(A, B) \in E$ et $(B, C) \in E$. Si $F_{\chi^2(2)}(\chi^2_{AC|B}) > F_{\chi^2(1)}(\chi^2_{AC})$ alors on supposera que $\{ABC\}$ forment une V-structure ($A \rightarrow B \leftarrow C$). Dans le cas contraire, on supposera que la connexion est en série ($A \rightarrow B \rightarrow C$, $A \leftarrow B \leftarrow C$) ou qu'elle est divergente $A \leftarrow B \rightarrow C$.*

La Figure 3 montre le graphe à cette étape. Les arêtes potentielles sont également concernées par les V-structures. On observe que seuls les noeuds D et O constituent les centres des connexions convergentes. Ce sont du reste les seules V-structures du réseau *Asia* original.

3.4 Suppression des arêtes superflues

Dans cette dernière étape, on traite les arêtes dites *potentielles* (cf Chap. 3.2) lors de la construction du graphe non orienté. On connaît désormais les chemins entre chaque paire de noeuds. On peut donc raisonner ainsi : pour chaque arête (AB) dans le graphe tel qu'il existe un autre chemin entre A et B sans V-structure, alors si conditionnellement à un caractère quelconque sur ce chemin, A et B deviennent indépendants, alors l'arête (AB) n'a pas lieu d'être. Cette proposition, très similaire à l'étape de "'Thinning"' dans la procédure *BN-Power Constructor* de Cheng et al. (2002), se généralise à tous les chemins entre A et B sans V-structures.

Proposition 3 *Soit $G(V, E)$ un graphe causal et $(A, B) \in E$. Posons $chemin_i(A, B)$, les chemins entre A et B uniquement formés de connexions de type CCC (en série ou divergente pour chaque noeud du chemin). Si*

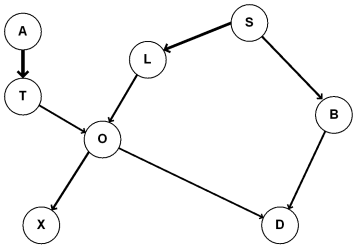

FIG. 1 – *Réseau Asia original. A désigne 'visite en Asie', T 'tuberculose', S 'fumeur', B 'bronchite', D 'dyspnée' (difficulté respiratoire), X 'rayons X' et L 'cancer de la langue'.*

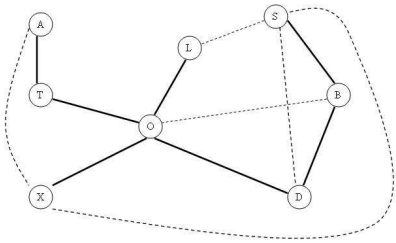

FIG. 2 – *Graphe non orienté résultant de l'algorithme 2. Les arêtes en pointillé sont les arêtes dites potentielles, les arêtes en trait plein sont les arêtes définitives.*

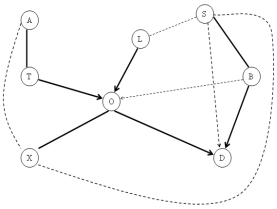

FIG. 3 – *Détection des V-structures par l'algorithme 3.*

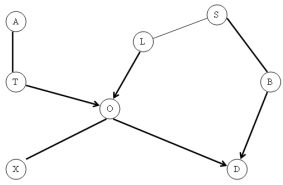

FIG. 4 – *Suppression de certaines arêtes potentielles par l'algorithme 4.*

$\chi^2(AB|S) < \chi^2_{indep}(2^{card(S)})$ avec $S = \cup S_i$ et S_i un noeud quelconque du $chemin_i(A, B)$, alors l'arête (A, B) doit être retirée de E.

L'algorithme 4 (non représenté) élimine les arêtes superflues du graphe contenant les V_structure en exploitant la proposition 3. Les chemins contenant une V_structure sont ignorés. Après suppression des arêtes superflues, le graphe partiellement dirigé obtenu par l'algorithme 4 est affiché sur la Figure 4. On observe que toutes les arêtes potentielles ont été éliminées hormis $L - S$. Il reste à propager les direction des arêtes en évitant l'ajout de V-structures. (A, T), (L, S),(S, B) peuvent être dirigées dans les deux sens, cependant, on ne doit pas créer la V-structure $L \rightarrow S \leftarrow B$. Le graphe obtenu est un *équivalent de Markov* du réseau original, il représente la même distribution de probabilités que le réseau *Asia*.

4 Application au cancer

4.1 Les données

Au vu des premiers résultats encourageants obtenus sur le réseau *Asia*, nous appliquons la méthode aux données d'une étude épidémiologique cas-témoins du cancer du nasopharynx (NPC). Pour clarifier le rôle de l'environnement dans l'étiologie du NPC, l'Unité d'épidémiologie génétique du CIRC a mené en 2004 une étude multicentrique de cas-témoins dans la

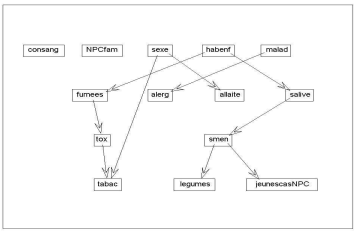

FIG. 5 – *RB obtenu après apprentissage de la structure sur une restriction des caractères susceptibles d'être impliqués dans le NPC chez les jeunes uniquement. La variable 'JeunesCasNPC' désigne l'occurrence du NPC chez les jeunes de moins de 25 ans.*

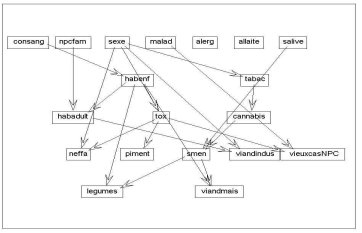

FIG. 6 – *RB obtenu après apprentissage de la structure sur une restriction des caractères susceptibles d'être impliqués dans le NPC chez les adultes uniquement. La variable 'VieuxCasNPC' désigne l'occurrence du NPC chez les personnes de plus de 35 ans.*

région endémique du Maghreb. 625 cas et 625 témoins ont été recrutés. Les données extraites du questionnaires comporte 1250 enregistrements et 480 caractères. A ce stade, nous faisons implictement deux hypothèses fortes : (1) l'efficacité de l'algorithme ne se dégrade pas trop lorsque l'on passe de 15000 exemples (*Asia*) à 1250 exemples ; (2) toutes les variables d'intérêt sont connues (principe de *suffisance causale*). Ce projet en phase préliminaire a pour but d'identifier les conjonctions de facteurs par des réseaux bayésiens, dans la lignée des travaux de recherche récents (Antal et al., 2004; Lucas et al., 2004; Getoor et al., 2004; Acid et al., 2004). L'idée est d'apporter un éclairage utile et pertinent sur les différents facteurs causes impliquées dans le NPC et dresser un profil statistique type de la population étudiée, que ne permet pas nécessairement la régression logistique, modèle couramment employée en épidémiologie.

4.2 Résultats

Dans cette section, nous présentons succinctement un travail - très préliminaire - mené sur un sous ensemble des données du questionnaire avec un petit réseau bayésien de quelques dizaines de méta variables booléennes, englobant plusieurs variables pré-sélectionnées, construites avec un expert à partir de la base de données. La variable à expliquer est " jeunescasNPC " dans la Figure 5, et " vieuxcasNPC " dans la Figure 6 (les épidémiologiste suspectent que les facteurs étiologiques seraient différents selon l'âge de survenue du cancer) ; les autres sont les variables explicatives : consanguinité, nombre de NPC dans la famille, sexe, habitat enfance, exposition aux fumées, maladies, allergies, exposition aux substances toxiques, allaitement, contact avec la salive adulte, tabac, smen (graisse animale), légumes, nourriture pimentée, neffa (drogue locale), cannabis, viande maison ou viande industrielle.

Les figurent 5 et 6 dressent les profils types des jeunes (<25 ans) et les individus plus âgés (>35 ans) atteints du NPC en mettant en évidence les causalités potentielles exhibées par l'algorithme. Il s'agit dans chaque cas d'une visualisation graphique des interactions entres les variables choisies et donc du profil statistique de la population considérées. Les tables de probabilités stockées dans les nœuds ne sont pas représentées. En observant uniquement

la structure du graphe, il est possible de dresser un profil moyen des populations analysées en gardant à l'esprit que seuls les arcs orientés lors de la découverte de V-structure (puis de la propagation de certains arcs après cette étape) correspondent à des relations causales. On observe que la cause principale du NPC chez les jeunes est la consommation de graisse animale saturée (smen) car cet arc ne peut être inversé sans créer de V-structure. En revanche, comme il n'y a aucune V-structure entre Smen, legume, et jeunesNPC, on a le choix en théorie entre plusieurs interprétations causales. L'expert du domaine opte pour habenf->salive->smen->legumes. Selon cette interprétation, le NPC chez les jeunes dépend indirectement les conditions d'habitat et le contact avec la salive adulte. A l'inverse, la consanguinité et l'allaitement n'ont manifestement aucune influence sur le risque de NPC. Chez les personnes plus âgées, c'est au contraire et les maladies et l'exposition aux produits toxiques qui semblent déterminant (V-structure), l'exposition étant liée aux conditions d'habitat à l'enfance. On retrouve également des résultats de bon sens : les hommes sont plus enclins à fumer, à consommer des drogues et être exposés à des produits toxiques (au travail), que l'exposition aux fumées (encens, parfums, feu de bois etc.) est plus fréquente dans les gourbis que les appartements en villes (habitat enfant) ; qu'il y a un lien entre l'habitat adulte et et le nombre de cas de cancer dans la famille ; que le smen se consomme avec de la viande maison et des légumes ; que les habitant des villes consomment plus de nourriture industrielle etc. On observe aussi que consanguinité, allergies et allaitement apparaissent comme variables indépendantes ; que certaines relations sont curieuses voire douteuses (e.g. le lien entre le nombre de cas familiaux et les conditions d'habitat à l'âge adulte) probalement en raison de l'existence de causes cachées (l'hypothèse de suffisance causale semble donc erronée).

En conclusion, ces graphes nous renseignent sur le mode de vie des sujets maghrébins et révèlent certaines chaînes de causalités susceptibles d'entraîner *in fine* le cancer de nasopharynx. Malgré le fait que de nombreux liens de causalité sont confirmés par les experts du domaine (e.g. smen, produits toxiques), nous nous gardons bien, à ce stade, de tirer des conclusions hâtives sur la pertinence des résultats obtenus.

5 Conclusion et critiques

Dans cet article, nous avons proposé une nouvelle méthode inspirée des travaux sur la recherche de motifs fréquents corrélés pour l'apprentissage de la structure des réseaux bayésien à partir de données. L'objectif est d'identifier les causalités entre les caractères à partir de tests d'indépendance conditionnelle et d'une mesure de corrélation, touts deux basés sur le χ^2. Après une rapide illustration sur le réseau $Asia$, nous avons appliqué la méthode aux données d'une étude épidémiologique cas-témoins du cancer du nasopharynx (NPC).

De nombreuses difficultés subsistent dans ce travail : le manque de fiabilité des tests d'indépendance conditionnelle dans les espaces de grande dimension (le nombre de données nécessaires augmentent exponentiellement en fonction du nombre de variables en condition), la nécessité de définir des seuils du $\chi 2$, un seuil pour le support, de garder en mémoire la nature des chemins (existence de V-structure ou non) entre chaque couple de variables. Ce travail présente encore de nombreuses faiblesses : il est impératif de mener une validation plus rigoureuse de la méthode sur les bancs d'essais disponibles sur Internet et d'étudier l'effet de la taille des bases d'exemples sur les performances. La complexité devra être examinée avec soin et les temps d'exécution comparée aux d'autres méthodes existantes (K2, GS, PC etc) dans l'esprit

des travaux de (Francois et Leray, 2004). Ces différents points font actuellement l'objet de nos travaux de recherche.

Références

Acid, S., L. M. de Campos, J. M. Fernández-Luna, S. Rodríguez, J. M. Rodríguez, et J. L. Salcedo (2004). A comparison of learning algorithms for bayesian networks : a case study based on data from an emergency medical service. *Artificial Intelligence in Medicine 30*(3), 215–232.

Adamo, J.-M. (2001). *Data Mining for asociation rules and sequential patterns*. Springer-Verlag New York.

Agrawal, R., T. Imielinski, et A. N. Swami (1993). Mining association rules between sets of items in large databases. In P. Buneman et S. Jajodia (Eds.), *SIGMOD conference, Washington, D.C.*, pp. 207–216. ACM Press.

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules in large databases. In J. B. Bocca, M. Jarke, et C. Zaniolo (Eds.), *International Conference on Very Large Data Bases (VLDB'94), Santiago de Chile, Chile*, pp. 487–499. Morgan Kaufmann.

Antal, P., G. Fannes, D. Timmerman, Y. Moreau, et B. D. Moor (2004). Using literature and data to learn bayesian networks as clinical models of ovarian tumors. *Artificial Intelligence in Medicine 30*(3), 257–281.

Bastide, Y., R. Taouil, N. Pasquier, G. Stumme, et L. Lakhal (2000). Mining frequent patterns with counting inference. *ACM SIGKDD Exploration 2*(2), 66–75.

Bayardo, R. J. (1998). Efficiently mining long patterns from databases. In L. M. Haas et A. Tiwary (Eds.), *ACM SIGMOD Conference, Seattle, USA*, pp. 85–93.

Bayardo, R. J. et R. Agrawal (1999). Mining the most interesting rules. In *international conference on Knowledge discovery and data mining (KDD'99)*, pp. 145–154.

Cheng, J., R. Greiner, J. Kelly, D. A. Bell, et W. Liu (2002). Learning bayesian networks from data : An information-theory based approach. *Artif. Intell. 137*(1-2), 43–90.

Cooper, G. (1997). A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery 1*(1), 203–224.

Francois, O. et P. Leray (2004). Evaluation d'algorithmes d'apprentissage de structure pour les réseaux bayésiens. In *Proceedings of 14ème Congrès Francophone Reconnaissance des Formes et Intelligence Artificielle, RFIA 2004*, Toulouse, France, pp. 1453–1460.

Friedman, N., K. Murphy, et S. Russell (1998). Learning the structure of dynamic probabilistic networks. In G. F. Cooper et S. Moral (Eds.), *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98)*, San Francisco, pp. 139–147. Morgan Kaufmann.

Getoor, L., J. T. Rhee, D. Koller, et P. Small (2004). Understanding tuberculosis epidemiology using structured statistical models. *Artificial Intelligence in Medicine 30*(3), 233–256.

Gunopulos, D., R. Khardon, H. Mannila, S. Saluja, H. Toivonen, et R. S. Sharma (2003). Discovering all most specific sentences. *ACM Transaction on Database System 28*(2), 140–174.

Han, J. et M. Kamber (2000). *Data Mining : Concepts and Techniques*. Morgan Kaufmann.

Han, J., J. Pei, et Y. Yin (2000). Mining frequent patterns without candidate generation. In W. Chen, J. F. Naughton, et P. A. Bernstein (Eds.), *International Conference on Management of Data (SIGMOD'00), Dallas, Texas, USA*, pp. 1–12. ACM.

Hand, D., H. Mannila, et P. Smyth (2001). *Principles of Data Mining*. MIT Press.

Lauritzen, S. et D. Speigelhalter (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Royal statistical Society B 50*, 157–224.

Leray, P. et O. Francois (2004). BNT structure learning package : Documentation and experiments. Technical report, Laboratoire PSI.

Lucas, P. J. F., L. C. van der Gaag, et A. Abu-Hanna (2004). Bayesian networks in biomedicine and health-care. *Artificial Intelligence in Medicine 30*(3), 201–214.

Murphy, K. (2001). The bayesnet toolbox for matlab. In *Computing Science and Statistics : Proceedings of Interface*, Volume 33.

Pasquier, N., Y. Bastide, R. Taouil, et L. Lakhal (1999). Efficient mining of association rules using closed itemset lattices. *Information Systems 24*(1), 25–46.

Pearl, J. (2000). *Causality : Models, Reasoning, and Inference*. Cambridge, England : Cambridge University Press.

Pearl, J. et T. S. Verma (1991). A theory of inferred causation. In J. F. Allen, R. Fikes, et E. Sandewall (Eds.), *KR'91 : Principles of Knowledge Representation and Reasoning*, San Mateo, California, pp. 441–452. Morgan Kaufmann.

Silverstein, C., S. Brin, R. Motwani, et J. D. Ullman (1998). Beyond market baskets : Generalizing association rules to correlations. In *VLDB*, pp. 594–605.

Silverstein, C., S. Brin, R. Motwani, et J. D. Ullman (2000). Scalable techniques for mining causal structures. *Data Min. Knowl. Discov. 4*(2/3), 163–192.

Spirtes, P., C. Glymour, et R. Scheines (2000). *Causation, Prediction, and Search* (2 ed.). The MIT Press.

Summary

Learning the structure of a bayesian network from a data set is NP-hard . Several methods have been proposed. In this paper, we discuss a new heuristic based on ideas developed for mining frequent sets in order to identify causal relations between random variables. We propose measuring significance of variable associations via the chi-square statistics. This leads to a measure that is upward closed in the itemset lattice, enabling us to reduce the mining problem to the search for a border between correlated and uncorrelated sets. A causal graph is then constructed from the correleted sets and the edges are afterward directed according to the V-structures found in the graph. A few unnecessary edges are then detected by subsequent idependence tests and removed. The method is first successfully applied on the Asia network benchmark to recover the original structure from data. The algorithm is then applied on the nasopharyngeal cancer (NPC) data in order to identify the environmental factors of the NPC.