Réseaux Bayésiens pour la Classification Méthodologie et Illustration dans le cadre du Diagnostic Médical

Philippe Leray — Olivier François

INSA Rouen / PSI, FRE CNRS 2645
BP 08 - Av. de l'Université
76801 St-Etienne du Rouvray Cedex
Philippe.Leray@insa-rouen.fr,
Olivier.Francois@insa-rouen.fr

RÉSUMÉ. *Les réseaux bayésiens sont des outils privilégiés pour les problèmes de diagnostic. Nous dressons dans cet article un panorama des algorithmes utilisés classiquement pour la mise en œuvre des réseaux bayésiens dans le cadre du diagnostic, et plus particulièrement du diagnostic médical. Pour cela, nous passons en revue un certain nombre de questions méthodologiques concernant le choix de la représentation des densités de probabilité (faut-il discrétiser les variables continues ? utiliser un modèle gaussien ?) et surtout la détermination de la structure du réseau bayésien (faut-il utiliser un réseau naïf ou essayer d'apprendre une meilleure structure à l'aide d'un expert ou de données ?). Une étude de cas concernant le diagnostic de cancer de la thyroïde nous permettra d'illustrer une partie de ces interrogations et des solutions proposées.*

ABSTRACT. *Bayesian networks are well suited tools for diagnosis tasks. In this paper, we focus on classical algorithms used to build diagnosis systems based on bayesian networks, and more particularly, medical diagnosis systems. We review some methodological questions concerning the representation of probability densities (discretization ? use of gaussian models ?) and the choice of the adequate structure (naive Bayes structure ? learning the structure with the help of an expert or from data ?). A case study, thyroid cancer diagnosis, will illustrate those considerations and some implemented algorithms*

MOTS-CLÉS : *diagnostic médical, apprentissage de paramètres, apprentissage de structure*

KEYWORDS: *medical diagnosis, parameter learning, structure learning*

RIA – 15/2002. Réseaux Bayésiens, pages 1 à 25

1. Introduction

Un diagnostic médical est le résultat du raisonnement d'un médecin, décision très souvent prise à partir d'informations incertaines et/ou incomplètes. De nombreuses techniques d'intelligence artificielle ont été appliquées pour essayer de modéliser ce raisonnement [LAV 97, LAV 99]. Ainsi, [SZO 82] présente l'utilisation détaillée de plusieurs systèmes experts en médecine. Citons, par exemple, des systèmes à base de règles comme MYCIN [SHO 74, BUC 84] et Internist-1/QMR (Quick Medical Reference) [MIL 82].

En amont de ce raisonnement, il faut aussi être capable de modéliser ces informations incertaines et/ou incomplètes. Certaines approches ont utilisé des formalismes comme la logique floue ([STE 97]) ou les fonctions de croyance de Dempster-Shafer. Une autre consiste à se placer dans le cadre de la théorie des probabilités, ce qui nous amène tout naturellement aux réseaux bayésiens (RB) proposés par Pearl [PEA 88] dans les années 80, retrouvés parfois sous le nom de systèmes experts probabilistes.

L'utilisation des réseaux bayésiens pose un certain nombre de questions méthodologiques :

– comment choisir la structure du RB ?

– comment représenter les densités de probabilités des variables continues ?

– comment estimer les densités de probabilités ?

– comment prendre en compte les données incomplètes ou les variables latentes ?

– comment faire de l'inférence, i.e. calculer la probabilité de telle ou telle maladie sachant certains symptômes ? , ...

Le but de cet article n'est pas d'exposer une méthode "révolutionnaire" d'aide au diagnostic médical, ni de répondre de manière exhaustive à toutes ces questions. Nous nous proposons de passer en revue la plupart des solutions qu'il est possible de mettre en œuvre, en illustrant certaines de ces techniques sur une étude de cas, un problème de diagnostic de cancer de la thyroïde.

2. Réseaux Bayésiens et Diagnostic Médical

2.1. *Quelques questions méthodologiques*

Les réseaux bayésiens possèdent de nombreux avantages (modélisation probabiliste de l'incertitude, possibilité de raisonnement aussi bien dans le sens symptômes-diagnostic que dans le sens diagnostic-symptômes, ...) qui font d'eux des outils privilégiés dans le cadre du diagnostic, notamment pour des problèmes de diagnostic médical où ils ont été utilisés dès les années 80 (cf. [KAP 00, SIE 00] pour une présentation de quelques applications de RB dans le domaine médical).

La mise en œuvre d'un RB pour modéliser un tel problème est assez immédiate lorsque celui-ci est simple (peu de variables, suffisamment de données et/ou dispo-

nibilité d'un expert pour l'apprentissage des probabilités). Ainsi, le *classifieur naïf de Bayes*, utilisé depuis longtemps en reconnaissance des formes statistiques, peut être vu comme un réseau bayésien très simple dont toutes les variables sont discrètes, avec l'hypothèse que tous les symptômes sont indépendants conditionnellement au diagnostic. Mais se pose alors une question classique dans la communauté *Machine Learning* : comment *discrétiser les variables continues* ?

Ce RB naïf peut bénéficier des apports de la communauté "réseaux bayésiens" pour contourner cette difficulté, en faisant l'hypothèse que la densité de probabilité conditionnelle (CPD) est une gaussienne (*RB naïf mixte*), ou un mélange de gaussiennes.

Un des inconvénients des RB naïfs est le nombre élevé de paramètres à estimer alors que, dans la plupart des cas, le nombre de données disponibles est faible. Pour y faire face,il est possible de modéliser les CPD par une fonction de type *OU bruité*. C'est ainsi que QMR/DT, une des premières applications de ce type de modélisation à un problème de diagnostic médical, a donné son nom par extension à ce type de RB (souvent appelé directement *QMR*)

Les RB naïfs ou de type QMR ont tous deux une structure simple à deux niveaux avec d'un côté les symptômes, et de l'autre les diagnostics. Dans la plupart des cas, le problème à résoudre est plus complexe à modéliser et la connaissance de certaines relations de causalité permet de construire un RB moins "naïf". Cette structure peut être obtenue grâce à un expert du domaine, ou à partir de données grâce à des méthodes d'*apprentissage de structure*.

Pour finir, il est aussi possible de modéliser des tâches de diagnostic encore plus complexes, en utilisant des architectures mixtes (réseaux de neurones, arbres de décision, réseaux bayésiens, ...), les RB étant utilisés au même niveau que les autres méthodes de classification, ou pour combiner efficacement les résultats des classifieurs. Nous ne décrirons pas ces méthodes ici, mais nous conseillons la lecture de [SIE 01] pour l'utilisation d'un RB pour la fusion de classifieurs pour le diagnostic médical, et de [LER 98] pour une illustration d'un système de diagnostic complexe (non médical), utilisant des réseaux de neurones (pour la reconnaissance de symptômes à partir de données brutes et pour la prise en compte de l'évolution temporelle) puis un réseau bayésien (pour le diagnostic final).

Après la phase de définition de la structure et du type des variables (discrètes, continues à CPD gaussiennes), il reste encore deux problèmes à résoudre. Tout d'abord, comment *estimer les probabilités conditionnelles* correspondant à la structure du RB (si ce n'est pas effectué en même temps que l'apprentissage de structure) ? Ensuite, la dernière question l'inférence, i.e. le calcul de la probabilité d'un (ou plusieurs) nœud(s) du RB (généralement, la variable *diagnostic*) conditionnellement à un ensemble d'observations. Un certain nombre d'algorithmes d'*inférence "exacte"* fonctionnent efficacement pour la plupart des RB. Par contre, dans certains cas, le réseau est trop complexe pour ces algorithmes, et il faudra utiliser des algorithmes d'*inférence "approchée"*.

2.2. *Choix de la structure du RB*

2.2.1. *RB Naïf*

Le classifieur de Bayes naïf est directement issu de l'application de la règle de décision de Bayes en rajoutant l'hypothèse d'indépendance conditionnelle des symptômes ($\mathbf{X}$) conditionnellement au diagnostic ($Diag$) :

$$\begin{aligned} d^o(\mathbf{X}) &= argmax_{Diag}\ p(Diag|\mathbf{X}) = argmax_{Diag}\ p(\mathbf{X}|Diag)p(Diag) \\ &= argmax_{Diag}\ \prod_i p(X_i|Diag)p(Diag) \end{aligned} \quad [1]$$

Cela nous permet de réécrire la loi jointe de la façon suivante, ce qui correspond graphiquement à la structure de la figure 2 p.13, appliquée à un problème de détection de cancer de la thyroïde.

$$p(\mathbf{X}, Diag) = p(Diag) \prod_i p(X_i|Diag) \quad [2]$$

Les implémentations classiques du classifieur de Bayes naïf considèrent que toutes les variables sont discrètes. Si certaines variables sont continues, il faut alors passer par une première étape de discrétisation. Cette étape, classique dans bon nombre d'algorithmes de *Machine Learning*, a été abordée de nombreuses fois, en utilisant des critères basés sur des tests statistiques ou sur la théorie de l'information [DOU 95]. Parmi ces méthodes, citons celles basées sur le critère d'Akaïke utilisées par [El- 00] pour la détection de mélanomes par un *réseau bayésien naïf discret*.

Une autre solution consiste à utiliser la modélisation CG (Conditional Gaussian) [LAU 92]. Sous certaines conditions, il est possible de représenter les densités de probabilités conditionnelles (CPD) continues par des gaussiennes. Il est alors possible de remplacer l'étape de discrétisation du RB naïf discret par une hypothèse de normalité des probabilités des symptômes conditionnellement au diagnostic pour obtenir ce que nous appellerons un ***RB naïf mixte***. Cette hypothèse assez forte permet cependant de réduire le nombre de paramètres à estimer ensuite (une moyenne et une variance à la place d'un histogramme).

De même,on peut relâcher l'hypothèse de normalité en remplaçant la CPD gaussienne par un mélange de gaussiennes. Cela se fait très facilement en rajoutant une variable latente (i.e. jamais mesurée) discrète entre le diagnostic et chaque symptôme.

2.2.2. *Modélisation OU bruité*

Dans les problèmes de diagnostic, la CPD importante à estimer est :

$$P = p(Diag|\mathbf{X}) = p(Diag|X_1, X_2, ..., X_n) \quad [3]$$

Supposons que la variable $Diag$ et les symptômes X_i soient binaires, de valeurs respectives $\{d$ et $\bar{d}\}$ et $\{x_i$ et $\bar{x}_i\}$. Pour estimer P, il faudra alors estimer 2^n valeurs, ce

qui n'est pas réaliste en grande dimension et/ou avec peu de données. L'idée est alors de simplifier cette probabilité en faisant les hypothèses suivantes :

– il est possible de calculer facilement la probabilité suivante (probabilité que X_i cause $Diag$ lorsque les autres variables X_j sont absentes) :

$$p_i = p(d|\bar{x}_1, \bar{x}_2, ..., x_i, ..., \bar{x}_n) \quad [4]$$

– le fait que X_i cause $Diag$ est indépendant des autres variables X_j (pas d'effet mutuel des variables).

Le modèle *OU bruité* (*noisy-OR*) permet d'estimer P par la formule suivante :

$$P = p(Diag|X_1, X_2, ..., X_n) = 1 - \prod_{i|X_i \in \mathbf{X}_p} (1 - p_i) \quad [5]$$

où $\mathbf{X}_p$ est l'ensemble des X_i vrais.

On peut remarquer que la nouvelle écriture de P ne fait pas d'hypothèses d'indépendance conditionnelle sur les X_i , ce qui correspond graphiquement à une structure de RB naïf où le sens de toutes les flèches aurait été inversé.

Ce modèle, proposé initialement par Pearl [PEA 86], a été étendu au cas où $Diag$ peut être vrai sans qu'un seul des symptômes soit présent ([HEN 89] *leaky noisy-OR gate*) et aux variables multivaluées ([HEN 89, SRI 93] *generalized noisy-OR gate*, [DIE 93] *noisy-MAX*).

Cette approche a donné de bons résultats dans de nombreux domaines. Dans le cadre du diagnostic médical, Shwe, Middleton et al. [SHW 91, MID 91] ont reformulé le système expert Internist/QMR sous la forme d'un réseau bayésien (QMR/DT) en utilisant le modèle OU bruité. [LEP 92] utilise le même type de modélisation pour un problème de détection de complications au cours de transfusions sanguines. Leur réseau bayésien possède 16 nœuds : 10 signes cliniques ou biologiques et 6 complications susceptibles de se déclencher.

Dans le cadre du diagnostic de problèmes hépatiques, [ONI 00] utilise un RB de 73 nœuds (66 caractéristiques et 7 diagnostics), où 27 des 73 CPD sont représentées par des OU bruités.

CPCS-PM (Computer-based Patient Case Simulation system), autre extension de Internist-1, a donné lieu lui aussi à l'utilisation de RB [PRA 94] avec une modélisation de type *noisy-MAX* à plusieurs niveaux (utilisation de variables intermédiaires entre les symptômes et les diagnostics) pour obtenir un RB de 448 nœuds et 908 arcs.

2.2.3. *Apprentissage de la structure*

Généralités

Comment trouver la structure qui représentera le mieux notre problème ? Dans le cas où les données sont complètes et décrivent totalement le problème (pas de variables latentes), la première étape est de mesurer l'adéquation d'un réseau bayésien à un

problème donné, d'*associer un score à chaque réseau bayésien*. La plupart des scores proposés dans la littérature sont décomposables en deux termes : le premier, la vraisemblance $p(\mathbf{D}|\theta, B)$, mesure l'adéquation du réseau bayésien de structure B et de paramètres θ aux données **D**. Le second terme va essayer de tenir compte de la complexité du modèle à l'aide, entre autres, du nombre de paramètres nécessaires pour représenter les distributions de probabilités du réseau (où r_i représente la taille de la variable X_i) :

$$Dim(B) = \sum_{X_i} (r_i - 1) \prod_{X_j \in pa(X_i)} r_j \qquad [6]$$

Parmi les différents scores proposés, citons les critères AIC [AKA 70] et BIC [SCH 78] dont les principes peuvent s'appliquer aux réseaux bayésiens :

$$ScoreAIC(B, \mathbf{D}) = \log p(\mathbf{D}|\theta^{MV}, B) - Dim(B) \qquad [7]$$

$$ScoreBIC(B, \mathbf{D}) = \log p(\mathbf{D}|\theta^{MV}, B) - \frac{1}{2} Dim(B) \log N \qquad [8]$$

où N est le nombre d'exemples dans **D** et θ^{MV} sont les paramètres obtenus par maximum de vraisemblance (cf. paragraphe 2.3).

On retrouve dans les équations 7 et 8 le principe du rasoir d'Occam : équilibrer la capacité à bien modéliser les données et à garder un modèle simple, repris dans les travaux sur la régularisation des réseaux de neurones [GIR 95].

Les autres scores existants sont soit des applications de mesures générales comme la longueur de description minimale MDL [BOU 93, SUZ 99], soit des mesures spécifiques aux réseaux bayésiens (Bayesian Mesure [COO 92], BDe [HEC 94], etc...).

La tâche suivante consiste à trouver le réseau qui donnera le meilleur score dans l'espace des RB. Une approche exhaustive est irréalisable en pratique, à cause de la taille de l'espace de recherche. Le nombre de structures possibles à partir de n variables, $NS(n)$, est donné par la formule de récurrence suivante [ROB 77], qui est super-exponentielle (par exemple, $NS(5) = 29281$ et $NS(10) = 4.2 \times 10^{18}$).

$$NS(n) = \begin{cases} 1 & , \quad n = 0 \; ou \; 1 \\ \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-1)} NS(n-i), & n > 1 \end{cases} \qquad [9]$$

Pour résoudre ce problème, un certain nombre d'heuristiques ont été proposées pour *parcourir l'espace des RB*.

Arbre de recouvrement minimal

Il est tout d'abord possible de se limiter à l'espace (beaucoup plus pauvre) des arbres. Une méthode dérivée de la recherche de l'arbre de recouvrement de poids minimal (*minimum weight spanning tree* ou MWST) a été proposée par [CHO 68]. Elle peut s'appliquer directement à la recherche de structure d'un réseau bayésien en fixant un poids à chaque arête potentielle X_i–X_j de l'arbre, par exemple l'*information mutuelle* entre les variables X_i et X_j tel que l'a présenté [CHO 68], ou encore la variation

du score lorsqu'on choisit X_j comme parent de X_i ([HEC 94]). L'arbre non dirigé retourné par les algorithmes classiques tels que Kruskal ou Prim doit ensuite être dirigé en choisissant arbitrairement un nœud racine puis en parcourant et orientant l'arbre par une recherche en profondeur.

Réseau bayésien naïf augmenté
Il est possible d'allier la simplicité du réseau bayésien naïf avec la puissance descriptive d'un RB plus général en rajoutant des dépendances directes entre les variables (indépendantes conditionnellement à la classe dans le modèle naïf). Parmi les méthodes simples pour *augmenter* le réseau naïf, citons le *Tree Augmented Naive Bayes* [KEO 99, FRI 97] qui s'obtient en cherchant le meilleur arbre reliant les observations (par l'algorithme MWST), puis en reliant toutes les observations à la classe comme pour un RB naïf classique [GEI 92]. [SAC 02] utilise différents classifieurs de type naïf augmenté pour l'interprétation d'images cardiaques SPECT.

Ordonnancement des nœuds
D'autres méthodes limitent l'espace de recherche en fixant un ordre de parcours des nœuds, puis en cherchant la meilleure configuration possible de parents pour chaque nœud parmi les nœuds suivants de la liste. Parmi ces méthodes, citons celle de référence, K2 (avec l'utilisation du score Bayesian Mesure) [COO 92] et des variantes comme K3 [BOU 93] (avec un score MDL), SGO [JOU 00] (avec une heuristique supplémentaire parcourant les énumérations possibles).

[WU 01] propose d'utiliser un RB pour la prédiction de survie en cas d'accident grave. Leur problème est assez représentatif des problèmes de diagnostic médical : peu de données (326 exemples) avec un nombre important de variables (29) et des données incomplètes. Dans cette approche, les auteurs commencent tout d'abord par un RB construit par un expert du domaine, puis par un RB construit par un algorithme proche de K2 prenant en compte les données manquantes. Ils utilisent ensuite les connaissances de l'expert pour déterminer une série de contraintes simples sur l'ordonnancement des nœuds (ordonnancement nécessaire à K2) et obtiennent alors un troisième réseau plus intéressant que les deux premiers.

Recherche gloutonne et algorithmes génétiques
D'autres méthodes d'apprentissage de structure présentent une série d'opérateurs (ajout d'arc, suppression, inversion) et effectuent une recherche gloutonne (*greedy search* [CHI 95a]) avec l'aide éventuelle de certaines heuristiques pour faciliter la recherche (algorithmes SG et SG+ [JOU 00]), ou utilisent des algorithmes génétiques [LAR 96].

[SIE 98] développe un système de prédiction de survie (à 1, 3 et 5 ans) après détection d'un mélanome malin (cancer de la peau) en utilisant un apprentissage de structure basé sur les algorithmes génétiques. Ce RB possède 6 nœuds (5 variables et un diagnostic) et les données mesurées sur 8 ans contiennent 311 exemples. Ce système obtient de meilleurs résultats qu'un classifieur de Bayes naïf. Il faut noter que les auteurs concluent sur l'importance d'incorporer à ces méthodes de construction automatique des connaissances d'experts sur la structure à obtenir.

Recherche dans l'espace des équivalents de Markov
En partant du fait que plusieurs structures encodent la même loi de probabilité (équivalence de Markov) et possèdent alors le même score, d'autres méthodes d'apprentissage de structure plus récentes suggèrent de *parcourir l'espace des équivalents de Markov*, espace légèrement plus petit (mais toujours super-exponentiel) mais possédant de meilleures propriétés : [CAU 00, MUN 01, AUV 02], GES (greedy equivalence search) [CHI 95b, CHI 96, CHI 02].

Recherche de causalité
Toutes ces méthodes font l'hypothèse de suffisance causale : toutes les variables d'intérêt sont connues. Pourtant, dans de nombreux cas, deux variables jugées dépendantes ne le sont que par des dépendances cachées (causes ou conséquences d'une troisième variable jamais mesurée). Ce problème a été étudié par certaines méthodes d'apprentissage de structure qui se concentrent sur la *notion de causalité* entre les variables plutôt que sur des scores de réseaux bayésiens. Deux séries d'algorithmes ont été proposées par deux équipes différentes : Pearl et Verma d'un côté avec les algorithmes IC et IC* (Inductive Causation) [PEA 91, PEA 00], Spirtes, Glymour et Scheines de l'autre avec les algorithmes SGS,PC, CI, FCI [SPI 93, SPI 00]. Ces algorithmes commencent tous par construire un graphe non dirigé contenant les relations entre les variables (à partir de tests d'indépendance conditionnelle) puis essaient de détecter les V-structures existantes (en utilisant aussi des tests d'indépendance conditionnelle). Il faut ensuite "propager" les orientations de certains arcs, et prendre éventuellement en compte les causes (et conséquences) artificielles dues à des variables latentes (algorithmes IC*, CI, FCI). Le principal inconvénient de ces méthodes de recherche de causalité est l'utilisation du test statistique d'indépendance conditionnelle qui donne des résultats peu fiables en grande dimension.

Traitement des données manquantes
Afin de compléter ce panorama des méthodes d'apprentissage de structure, citons enfin les méthodes EM structurelles [FRI 98] qui appliquent l'algorithme EM (décrit en 2.3 dans le cas de l'apprentissage de paramètres) à une recherche de structure de type gloutonne.

Indépendamment de la méthode utilisée, il semble assez illusoire de chercher la meilleure structure sans utiliser de connaissances a priori sur le problème à résoudre. Il est souvent possible de déterminer des sous-problèmes qui seront modélisés séparément, de définir par avance des groupes de variables qui sont liées, etc ... Ces connaissances fournies par des experts du domaine permettent de limiter fortement l'espace de recherche.

2.3. *Apprentissage des paramètres*

Après avoir trouvé la structure du réseau bayésien (ou pendant l'apprentissage de structure, selon les méthodes), il est nécessaire d'estimer les distributions de probabilités conditionnelles du réseau (ou les paramètres des lois correspondantes). Comme

pour tout problème d'apprentissage, différentes techniques sont possibles selon la disponibilité de données pour le problème à traiter, ou d'experts du domaine. On peut classer ces techniques en deux grandes familles : *apprentissage à partir de données* (complètes ou non), par des approches statistiques classiques ou bayésiennes, et *acquisition de connaissances* (avec un expert du domaine). Nous nous restreindrons ici aux RB à variables discrètes, les principes évoqués pouvant se généraliser aux RB conditionnels gaussiens ([LAU 92]).

2.3.1. *Apprentissage à partir de données*

L'estimation de distributions de probabilités (paramétriques on non) à partir de données est un sujet très vaste et complexe. Nous décrirons ici les méthodes les plus utilisées dans le cadre des réseaux bayésiens, selon que les données à notre disposition sont complètes ou non, en conseillant la lecture de [HEC 98, KRA 98, JOR 98a] pour plus d'informations.

Dans le cas où toutes les variables sont observées, la méthode la plus simple et la plus utilisée est *l'estimation statistique*. Il s'agit d'estimer la probabilité d'un événement par la fréquence d'apparition de l'événement dans la base de données. Cette approche (appelée *maximum de vraisemblance (MV)*) nous donne alors :

$$\hat{p}(X_i = x_k | pa(X_i) = x_j) = \hat{\theta}^{MV}_{i,j,k} = \frac{N_{i,j,k}}{\sum_k N_{i,j,k}} \qquad [10]$$

où $N_{i,j,k}$ est le nombre d'événements dans la base de données pour lesquels la variable X_i est dans l'état x_k et ses parents sont dans la configuration x_j.

Le principe, quelque peu différent, de *l'estimation bayésienne* consiste à trouver les paramètres θ les plus probables *sachant que les données ont été observées*. En utilisant une distribution de Dirichlet comme a priori sur les paramètres, on peut écrire :

$$p(\theta) = \prod_{i=1}^{n} \prod_{j} \prod_{k=1}^{r} (\theta_{i,j,k})^{\alpha_{i,j,k}} \qquad [11]$$

où $\alpha_{i,j,k}$ sont les paramètres de la distribution de Dirichlet associée à la loi a priori $p(X_i = x_k | pa(X_i) = x_j)$.

L'approche de *maximum a posteriori (MAP)* nous donne :

$$\hat{p}(X_i = x_k | pa(X_i) = x_j) = \hat{\theta}^{MAP}_{i,j,k} = \frac{N_{i,j,k} + \alpha_{i,j,k} - 1}{\sum_k (N_{i,j,k} + \alpha_{i,j,k} - 1)} \qquad [12]$$

Dans la plupart des applications, les bases d'exemples sont très souvent incomplètes. Certaines variables ne sont observées que partiellement ou même jamais. La méthode d'estimation de paramètres avec des données incomplètes la plus couramment utilisée est fondée sur l'algorithme itératif *Expectation-Maximisation* (EM) proposé par Dempster [DEM 77] et appliqué aux RB dans [COW 99, NEA 98].

Même si nous ne présentons ci-dessous que l'utilisation de l'algorithme EM aux réseaux bayésiens discrets, notons que le principe s'applique sans problème aux réseaux bayésiens de type *conditionnel gaussien* où certains nœuds sont continus et modélisés par des densités de probabilités conditionnelles gaussiennes. Cette problématique est d'ailleurs similaire à celle de l'apprentissage des mélanges de gaussiennes ou des modèles de Markov cachés [NEA 98, VLA 02].

Soient

– $\mathbf{X}_v = \{\mathbf{X}_v^{(l)}\}_{l=1...N}$ l'ensemble des données observées (visibles),

– $\theta^{(t)} = \{\theta_{i,j,k}^{(t)}\}$ les paramètres du réseau bayésien à l'itération t.

L'algorithme EM s'applique à la recherche des paramètres en répétant jusqu'à convergence les deux étapes *Espérance et Maximisation* décrites ci-dessous.

– **Espérance** : estimation des $N_{i,j,k}$ manquants en calculant leur moyenne conditionnellement aux données et aux paramètres courants du réseau :

$$N_{i,j,k}^* = E[N_{i,j,k}] = \sum_{l=1}^{N} p(X_i = x_k | pa(X_i) = x_j, \mathbf{X}_v^{(l)}, \theta^{(t)}) \qquad [13]$$

Cette étape revient à faire une série d'inférences (exactes ou approchées) en utilisant les paramètres courants du réseau, et à remplacer les valeurs manquantes par les probabilités obtenues par inférence.

– **Maximisation** : en remplaçant les $N_{i,j,k}$ manquants par leur valeur moyenne calculée précédemment, il est maintenant possible de calculer de nouveaux paramètres $\theta^{(t+1)}$ par maximum de vraisemblance :

$$\theta_{i,j,k}^{(t+1)} = \frac{N_{i,j,k}^*}{\sum_k N_{i,j,k}^*} \qquad [14]$$

L'algorithme EM peut aussi s'appliquer dans le cadre bayésien. Pour l'apprentissage des paramètres, il suffit de remplacer le maximum de vraisemblance de l'étape M par un maximum à posteriori. Cela nous donne donc :

$$\theta_{i,j,k}^{(t+1)} = \frac{N_{i,j,k}^* + \alpha_{i,j,k} - 1}{\sum_k (N_{i,j,k}^* + \alpha_{i,j,k} - 1)} \qquad [15]$$

2.3.2. *Extraction de connaissances*

Il existe de nombreux travaux sur l'extraction de probabilités (cf. [REN 01]). Lorsqu'un expert doit déterminer tout un ensemble de probabilités, il faut tenir compte des biais éventuels parfois subconscients (un expert va souvent surestimer la probabilité de réussite d'un événement le concernant plus directement, etc ...). Il est possible de fournir à cet expert du domaine des outils reliant des notions qualitatives et quantitatives

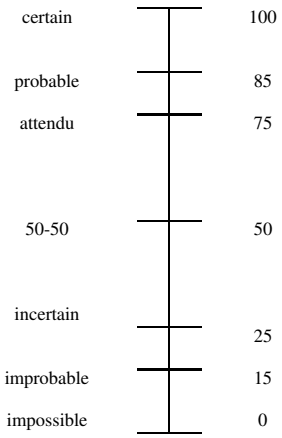

Figure 1. *Echelle de probabilité*

pour qu'il puisse associer une probabilité aux différents événements. L'outil le plus connu et le plus facile à mettre en œuvre est l'échelle de probabilité présentée dans la figure 1 (cf. les travaux de Druzdzel [DRU 00] et Renooij [REN 01]). Cette échelle permet aux experts d'utiliser des informations à la fois textuelles et numériques pour assigner un degré de réalisation à telle ou telle affirmation.

Une des applications les plus connues des réseaux bayésiens en médecine est le système Pathfinder [HEC 92], spécialisé dans le diagnostic des pathologies ganglionnaires. Cette application traite 130 symptômes et 60 diagnostics et nécessite la spécification d'environ 75.000 probabilités. Autre exemple, [GAA 02] étudie de façon très détaillée les techniques d'élicitation de probabilité pour la prédiction de l'état d'avancement d'un cancer de l'œsophage. Les auteurs ont à leur disposition 40 variables mesurées partiellement sur 156 exemples qu'ils préfèrent garder pour tester la validité des modèles obtenus. Deux spécialistes du domaine ont été interrogés pour déterminer la structure du RB et du millier de probabilités associées. Après une phase de réglage du RB et de correction de certaines données par les experts, le RB détermine correctement l'état du patient dans 85% des cas.

2.4. *Inférence*

L'inférence consiste à calculer la probabilité d'un (ou plusieurs) nœud(s) du réseau bayésien conditionnellement à un ensemble d'observations. Un certain nombre d'algorithmes permet, en théorie, de faire ce calcul de manière exacte. Nous conseillons la lecture de [PEA 88] et [JEN 96] pour une description des algorithmes d'inférence les plus couramment utilisés. Ces méthodes sont malheureusement trop lourdes à utiliser pour des réseaux de très grande taille, ou fortement connectés. Pour essayer de résoudre ces problèmes, des algorithmes d'inférence approchée ont été mis au point, par

exemple en utilisant des techniques d'échantillonage. D'autres méthodes approchées utilisent des approximations variationnelles développées récemment ([JOR 98b]).

[JAA 99] propose une approximation variationnelle des réseaux bayésiens de type QMR/DT. [WIE 99] utilise une méthode d'inférence variationnelle sur un projet de diagnostic de l'anémie (Promedas), avec un RB d'une centaine de variables. [KAP 02] décrit une méthode d'inférence variationnelle (Cluster Variation Method) qu'il applique avec succès au même problème.

3. Etude de cas : Cancer de la thyroïde

3.1. *Les données*

La base d'exemple utilisée est une base classique proposée par [QUI 86] dans le cadre des arbres de décision, et disponible sur de nombreux serveurs web. Elle est séparée en deux ensembles (apprentissage et test) contenant respectivement 2800 et 972 enregistrements. Parmi les 29 variables initiales, nous retenons ici l'ensemble des 22 variables décrit dans le tableau 1.

diag	état (0=sain et 1=malade)
X_1	âge (continue)
X_2	sexe (0=féminin et 1=masculin)
X_3	sous thyroxine
X_4	demande de thyroxine
X_5	sous traitement antithyroïde
X_6	malade
X_7	femme enceinte
X_8	opéré de la thyroïde
X_9	sous traitement I131
X_{10}	demande d'hypothyroïde
X_{11}	demande d'hyperthyroïde
X_{12}	sous lithium
X_{13}	présence d'un goitre
X_{14}	présence d'une tumeur
X_{15}	hypopituitaire
X_{16}	psych
X_{17}	mesure de TSH (continue)
X_{18}	mesure de T3 (continue)
X_{19}	mesure de TT4 (continue)
X_{20}	mesure de T4U (continue)
X_{21}	mesure de FTI (continue)

Tableau 1. *Thyroid : les 22 variables utilisées*

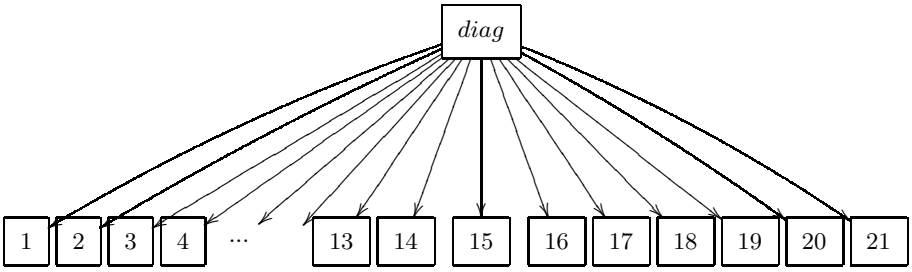

Figure 2. *Réseau bayésien naïf discret*

885	27	0
15	45	0

$\alpha = 0.5$ (pas de rejet)

836	11	65
6	27	27

$\alpha = 0.9$

Tableau 2. *Réseau bayésien naïf discret : matrices de confusion (en test) pour deux seuils de rejet. Les lignes correspondent à la classe réelle (sain, malade), les colonnes à la décision prise suivant les résultats du classifieur (sain, malade et rejet).*

3.2. *Réseau bayésien naïf discret*

Commençons par mettre en œuvre un RB naïf discret (fig. 2) en discrétisant les variables continues par une des méthodes proposées par [El- 00]. Les CPD sont estimées à partir des exemples d'apprentissage. Le RB naïf est ensuite utilisé pour calculer $p(Diag|\mathbf{X})$ et associé à une règle de décision avec rejet : si $max(p(Diag|\mathbf{X})) < \alpha$, alors décision $= rejet$, sinon décision $= argmax(p(Diag|\mathbf{X}))$. La table 2 nous donne les matrices de confusion correspondant à deux seuils de rejet.

Il est également possible d'évaluer la qualité du classifieur obtenu en traçant la courbe ROC (pourcentage d'exemples non rejetés bien classés en fonction du pourcentage des exemples rejetés). La figure 3 nous donne la courbe ROC du réseau naïf discret (courbe foncée en trait plein). Elle nous indique que, sans rejet, le réseau naïf discret donne 95.7% de bonne classification. De plus, pour être sur d'obtenir un pourcentage de bien classés de 99% il faudra rejeter 33.5% des exemples (et les traiter manuellement ou avec un autre classifieur).

3.3. *Réseau bayésien naïf mixte*

Remplaçons maintenant la discrétisation des variables pour une hypothèse supplémentaire (modélisation des CPD continues par des gaussiennes) pour obtenir le RB naïf mixte de la figure 4. Ce réseau, qui possède un nombre réduit de paramètres par rapport au RB naïf discret, nous donne le même pourcentage de bonne classification (95.7%), avec une meilleure courbe ROC (cf. figure 3, courbe grisée en trait plein). En

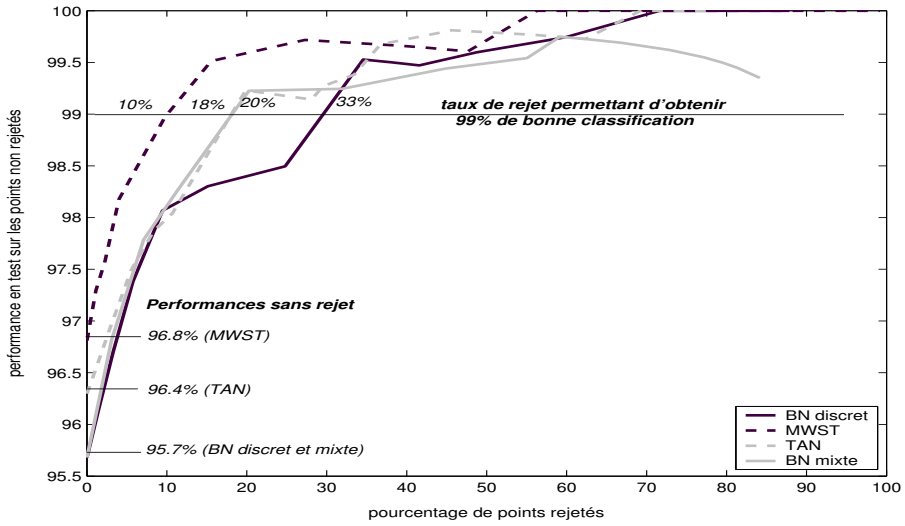

Figure 3. *Courbe ROC pour différents réseaux bayésiens (RB naïf discret, RB naïf mixte, RB obtenus par MWST et TAN).*

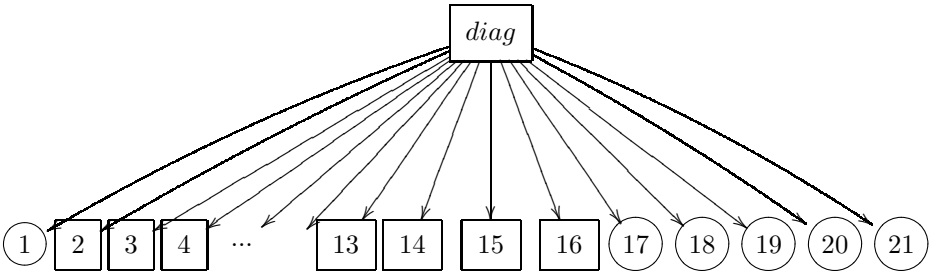

Figure 4. *Réseau bayésien naïf mixte. Les variables continues sont représentées par des ronds.*

effet, pour être sûr d'avoir un pourcentage de bien classés de 99% il faut maintenant rejeter seulement 18.2% des exemples (contre 33.5% pour le RB naïf discret).

3.4. *Arbre de recouvrement maximal et réseau naïf augmenté*

L'arbre de recouvrement maximal (MWST) [CHO 68] utilisé ici sur les données discrètes précédentes avec le score BIC de l'équation 8 nous donne l'arbre orienté de la figure 5. Malgré la restriction assez forte sur l'espace de recherche (passage de l'espace des graphes reliant les variables à l'espace des arbres), le réseau bayésien obtenu donne des résultats meilleurs que le réseau bayésien naïf discret : 96.8% de

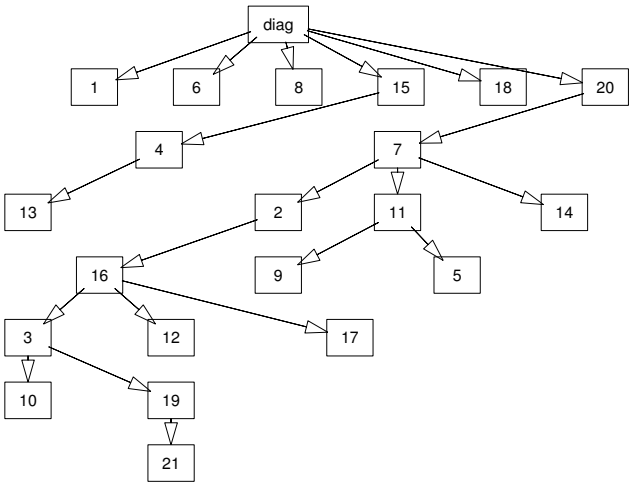

Figure 5. *Réseau bayésien obtenu par l'algorithme MWST (La racine choisie pour l'orientation de l'arbre est le nœud diagnostic).*

bonne classification sans rejet (contre 95.7% pour le naïf discret), et un pourcentage de bonne classification de 99% pour un taux de rejet de 10% (bien meilleur que pour les deux réseaux bayésiens naïfs, cf. la courbe foncée en pointillé de la figure 3).

Dans l'approche MWST, la connaissance a priori de la variable qui sert à la classification n'entre pas en jeu, à la différence de la structure proposée par le réseau bayésien naïf. L'approche TAN (Tree Augmented Naive bayes) permet de mélanger les deux, en cherchant le meilleur arbre reliant les observations et en conservant la structure reliant la classe aux observations. Le réseau ainsi obtenu donne des performances en test équivalentes (96.4%) mais avec des performances de rejet moins bonnes (20.5% de points rejetés pour arriver à 99% de bonne classification, cf. la courbe grisée en pointillé de la figure 3).

3.5. *Ordonnancement des nœuds, algorithme K2*

Cherchons maintenant si un RB de structure plus complexe pourrait mieux modéliser notre problème. N'ayant pas d'expert à notre disposition, nous allons appliquer l'algorithme K2 proposé par [COO 92]. Cet algorithme ne fonctionnant qu'avec des données discrètes, nous utiliserons donc les données déjà discrétisées en 3.2. En utilisant un ordonnancement des nœuds inspiré du RB naïf (d'abord le nœud Diagnostic, puis les autres nœuds), on obtient le RB de la figure 6 qui montre un pourcentage de bonne classification en test de 96.3%. Ce RB nous permet d'obtenir la courbe foncée en trait pointillé de la figure 8. Elle nous indique que pour être sûr d'avoir un pourcen-

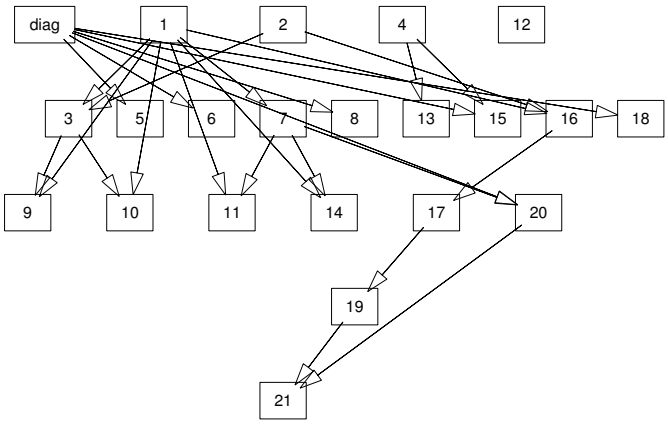

Figure 6. *Réseau bayésien obtenu par l'algorithme K2 avec l'ordre d'énumération* $Diag, X_1, X_2, ..., X_{21}$.

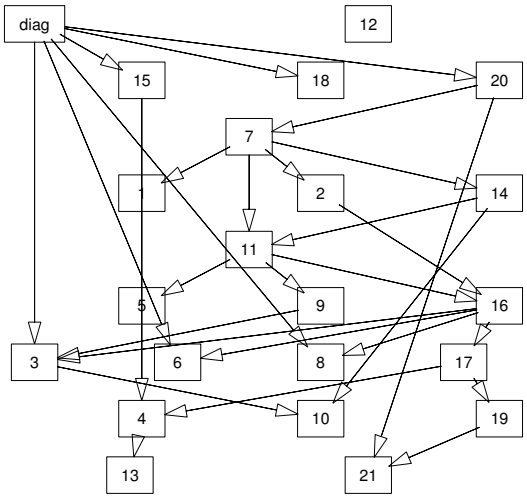

Figure 7. *Réseau bayésien obtenu par l'algorithme K2+T avec l'ordre d'énumération fourni par MWST.*

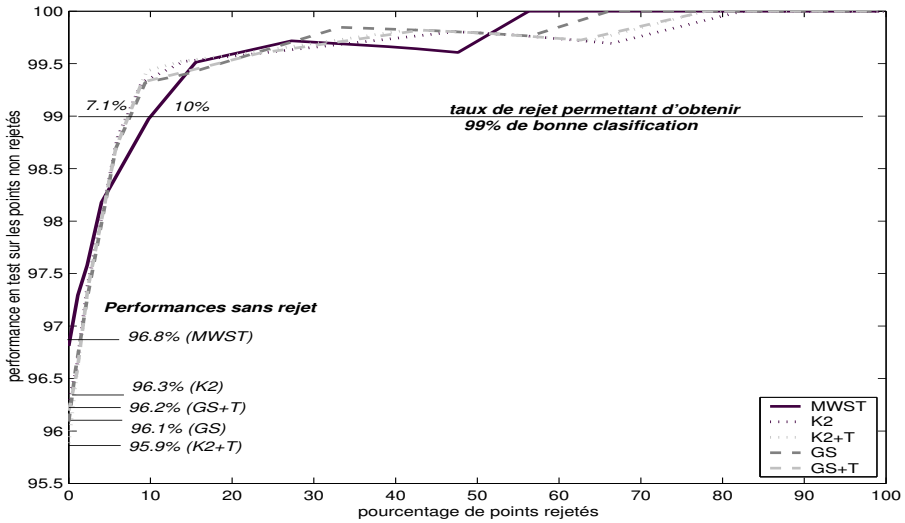

Figure 8. *Courbe ROC pour différents réseaux bayésiens (RB obtenus par MWST, K2, K2+T, GS et GS+T).*

tage de bien classés de 99% il faut maintenant rejeter 7.1% des exemples (contre 33% et 10% pour le RB naïf et pour l'arbre obtenu avec MWST).

Notons que le résultat de l'algorithme K2 dépend fortement de l'ordonnancement initial des nœuds. Un ordre différent aurait pu donner des résultats très mauvais ou éventuellement meilleurs ! Pour résoudre ce problème d'initialisation, nous avons proposé dans [FRA 03] d'utiliser l'ordonnancement des nœuds fourni par l'algorithme MWST pour initialiser l'algorithme K2. Cette variante de K2 appellée K2+T nous donne le réseau bayésien de la figure 7 et un pourcentage de bonne classification en test de 95.9%. La courbe ROC obtenue par K2+T est sensiblement la même que celle obtenue par K2 ; par conséquent nous avons réussi à obtenir un RB donnant des performances équivalentes, mais en nous affranchissant du problème d'initialisation.

3.6. *Recherche gloutonne, algorithme GS*

L'algorithme GS (recherche gloutonne, *Greedy Search*) permet de lever la restriction sur l'ordre des nœuds pour le parcours de l'espace des structures possibles. Une série d'opérateurs (ajout, suppression et inversion d'arc) définit le voisinage d'une structure fixée. Il suffit alors de rechercher une structure plus intéressante parmi le voisinage, et d'itérer la recherche jusqu'à convergence du critère de score [CHI 95a].

En partant d'une initialisation vide (structure sans arc), cette méthode nous donne le réseau bayésien de la figure 9. Les performances en classification sont équivalentes

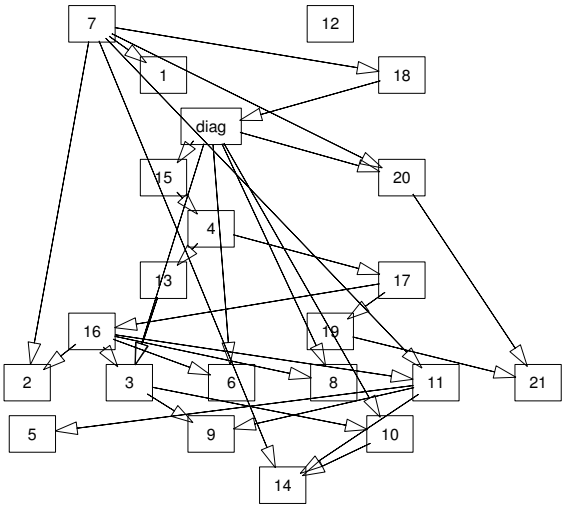

Figure 9. *Réseau bayésien obtenu par l'algorithme Greedy Search.*

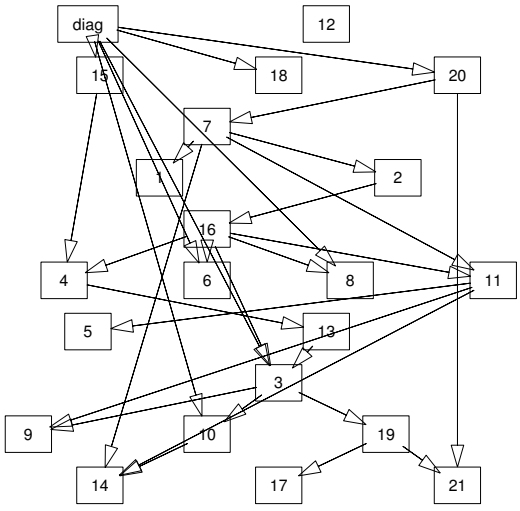

Figure 10. *Réseau bayésien obtenu par l'algorithme Greedy Search, avec une initialisation fournie par MWST.*

à celles des réseaux obtenus par K2 et K2+T (96.1% de bonne classification et un taux de rejet de 7.5% pour obtenir des performances en test de 99%).

La vitesse de convergence de ce genre de méthode dépend fortement de la structure utilisée au départ. Comme pour K2 et K2+T, nous avons proposé dans [FRA 03] d'utiliser cette fois-ci le graphe fourni par l'algorithme MWST pour initialiser l'algorithme GS. Cette variante appelée GS+T nous fournit, avec environ deux fois moins de calculs, le réseau bayésien de la figure 10 qui possède des performances du même ordre que les autres méthodes (96.2% de bonne classification et 7.2% de rejet pour obtenir des performances en test de 99%).

3.7. *Algorithme EM structurel*

Dans toutes les méthodes précédentes, le problème des données manquantes (variables partiellement observées) était contourné en rajoutant une modalité supplémentaire (*variable non mesurée*) aux variables concernées.

Une manière plus formelle de résoudre le problème est d'utiliser le principe de l'algorithme EM de [DEM 77] à l'apprentissage de structure. L'algorithme itératif SEM (*Structural EM*) proposé par [FRI 98] combine un algorithme de type *Greedy Search* pour définir le voisinage de la structure courante, et l'algorithme EM pour évaluer les paramètres et le score de tous les réseaux de ce voisinage, et choisir le meilleur pour l'itération suivante. Cet algorithme souffre encore de problèmes d'initialisation qui font qu'il est souvent utile de l'exécuter plusieurs fois pour éviter de tomber dans des minima locaux de très mauvaise qualité. Malgré cela, il est possible d'arriver à des solutions intéressantes obtenant un taux de bonne classification proche des autres méthodes.

4. Conclusion

Dans cet article, nous avons dressé un panorama d'algorithmes classiquement utilisés pour la mise en œuvre de réseaux bayésiens dans le cadre du diagnostic, et plus particulièrement du diagnostic médical. Pour aborder plus concrètement cette tâche, nous avons appliqué un certain nombre d'algorithmes sur un problème de détection du cancer de la thyroïde. Le tableau 3 résume les performances obtenues avec plusieurs méthodes d'apprentissage de structure, avec ou sans discrétisation des variables continues. Cette étude nous a permis d'aborder certaines questions méthodologiques simples mais qui se posent lors de toutes les applications :

– comment représenter les densités de probabilités des variables continues ? faut-il discrétiser ? représenter les CPD continues par des gaussiennes ?
L'utilisation d'une CPD gaussienne simple peut poser des problèmes si la distribution est bimodale, et l'utilisation de mélanges de gaussiennes pose d'autres difficultés comme la détermination du nombre de gaussiennes à utiliser. De plus, certaines méthodes d'apprentissage de structure ne peuvent s'utiliser qu'avec des variables dis-

Méthode	Perf. (sans rejet)	Intervalle de confiance	Rejet (/ Perf=99%)
BN discret	95.7%	[94.2% – 96.9%]	33.5%
BN mixte	95.7%	[94.2% – 96.9%]	18.2%
MWST	96.8%	[95.4% – 97.8%]	10%
TAN	96.4%	[95.0% – 97.5%]	20.5%
K2	96.3%	[94.9% – 97.4%]	7.1%
K2+T	95.9%	[94.4% – 97.0%]	7.1%
GS	96.1%	[94.6% – 97.2%]	7.5%
GS+T	96.2%	[94.7% – 97.3%]	7.3%

Tableau 3. *Thyroid : performances en test sans rejet (colonne 2) avec intervalle de confiance à 95% (colonne 3) et taux de rejet correspondant à 99% de bonne classification (colonne 4) pour des réseaux bayésiens obtenus par différents algorithmes d'apprentissage de structure.*

crètes. D'un autre côté, le nombre de paramètres à estimer est souvent plus petit dans le cas conditionnel gaussien, ce qui permet d'obtenir de meilleurs résultats.

– comment choisir la structure du RB ? faut-il utiliser un RB naïf, ou essayer de trouver une meilleure structure ?
L'utilitation d'un réseau bayésien naïf permet souvent d'obtenir de bons résultats à un moindre coût, mais est rapidement surclassée par MWST, méthode presque aussi simple. Par contre, si le nombre de données disponibles est important ou avec l'aide d'un expert, il est possible d'obtenir une structure codant plus finement le problème.

Les perspectives sont nombreuses, surtout au niveau de l'apprentissage de structure et plus spécifiquement l'apprentissage dans l'espace des équivalents de Markov et l'application de l'algorithme SEM dans le même espace. Il reste aussi à proposer des méthodes permettant d'incorporer automatiquement des connaissances a priori (méta-structures, connaissances d'experts, ...) pour faciliter la recherche de la structure et améliorer la convergence de méthodes comme la recherche gloutonne ou SEM. Une autre voie de recherche concerne les réseaux bayésiens temporels qui offrent un cadre idéal pour la prise en compte du temps dans le diagnostic. Pour finir, il pourrait être intéressant d'essayer de modéliser l'incertain avec un autre formalisme que les probabilités, en utilisant par exemple la théorie de Dempster-Schafer.

Remerciements

Les expérimentations effectuées dans cet article ont été réalisées avec BNT, toolbox gratuite pour Matlab [MUR 01] et le package *Structural Learning* que nous distribuons sur le site internet français de la toolbox (http ://bnt.insa-rouen.fr).

5. Bibliographie

[AKA 70] AKAIKE H., « Statistical Predictor Identification », *Ann. Inst. Statist. Math.*, vol. 22, 1970, p. 203-217.

[AUV 02] AUVRAY V., WEHENKEL L., « On the Construction of the Inclusion Boundary Neighbourhood for Markov Equivalence Classes of Bayesian Network Structures », DARWICHE A., FRIEDMAN N., Eds., *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI-02)*, S.F., Cal., 2002, Morgan Kaufmann Publishers, p. 26–35.

[BOU 93] BOUCKAERT R., « Probabilist network construction using the Minimum Description Length principle », rapport, 1993, Departement of computer science, Utrech university, Netherlands.

[BUC 84] BUCHANAN B., SHORTLIFFE E. H., *Rule-Based Expert Systems : The MYCIN Experiments of the Stanford Heuristic Programming Project*, Addison Wesley, 1984.

[CAU 00] CAU D., MUNTEANU P., « Efficient Learning of Equivalence Classes of Bayesian Networks », *Proceedings of the 4th European Conf. on Principles and Practice of Knowledge Discovery in Databases, PKDD, Lyon*, 2000, p. 96-105.

[CHI 95a] CHICKERING D., GEIGER D., HECKERMAN D., « Learning Bayesian networks : Search methods and experimental results », *Proceedings of Fifth Conference on Artificial Intelligence and Statistics*, 1995, p. 112–128.

[CHI 95b] CHICKERING D. M., « A Transformational Characterization of Equivalent Bayesian Network Structures », BESNARD, PHILIPPE, HANKS S., Eds., *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI'95)*, San Francisco, CA, USA, août 1995, Morgan Kaufmann Publishers, p. 87–98.

[CHI 96] CHICKERING D. M., « Learning Equivalence Classes of Bayesian Network Structures », HORVITZ E., JENSEN F., Eds., *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence (UAI-96)*, San Francisco, août 1–4 1996, Morgan Kaufmann Publishers, p. 150–157.

[CHI 02] CHICKERING D. M., « Learning equivalence classes of bayesian-network structures », *Journal of machine learning research*, vol. 2, 2002, p. 445-498.

[CHO 68] CHOW C., LIU C., « Approximating discrete probability distributions with dependence trees », *IEEE Transactions on Information Theory*, vol. 14, n° 3, 1968, p. 462-467.

[COO 92] COOPER.G, HERSOVITS.E, « A Bayesian Method for the Induction of Probabilistic Networks from Data », *Maching Learning*, vol. 9, 1992, p. 309-347.

[COW 99] COWELL R. G., DAWID A. P., LAURITZEN S. L., SPIEGELHALTER D. J., *Probabilistic Networks and Expert Systems*, Statistics for Engineering and Information Science, Springer-Verlag, 1999.

[DEM 77] DEMPSTER A., LAIRD N., RUBIN D., « Maximum Likelihood from Incompete Data Via the EM Algorithm », *journal of the Royal Statistical Society*, vol. B 39, 1977, p. 1-38.

[DIE 93] DIEZ F. J., « Parameter adjustement in Bayes networks. The generalized noisy OR–gate », *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence*, Washington D.C., 1993, Morgan Kaufmann, San Mateo, CA, p. 99–105.

[DOU 95] DOUGHERTY J., KOHAVI R., SAHAMI M., « Supervised and Unsupervised Discretization of Continuous Features », *International Conference on Machine Learning*, 1995, p. 194-202.

[DRU 00] DRUZDEL M., VAN DER GAAG L., HENRION M., JENSEN F., « Building Probabilistic Networks : "Where Do the Numbers Come From ?" Guest Editors Introduction », *IEEE Transactions on Knowledge and Data Engineering*, vol. 12, 2000.

[El- 00] EL-MATOUAT F., COLOT O., VANNOORENBERGHE P., LABICHE J., « From continous to discrete variables for bayésian network classifiers », *Conference on Systems, Man and Cybernetics, IEEE-SMC*, Nashville, USA, 2000.

[FRA 03] FRANCOIS O., LERAY P., « Etude comparative d'algorithmes d'apprentissage de structure dans les réseaux bayésiens », *Proceedings of RJCIA 2003, plateforme AFIA 2003*, Laval, France, 2003.

[FRI 97] FRIEDMAN N., GEIGER D., GOLDSZMIDT M., « Bayesian Network Classifiers », *Machine Learning*, vol. 29, n° 2-3, 1997, p. 131-163.

[FRI 98] FRIEDMAN N., « The Bayesian Structural EM Algorithm », COOPER G. F., MORAL S., Eds., *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98)*, San Francisco, juillet 24–26 1998, Morgan Kaufmann, p. 129–138.

[GAA 02] DER GAAG L. V., RENOOIJ S., WITTEMAN C., ALEMAN B., TAAL B., « Probabilities for a probabilistic network : a case study in oesophageal cancer », *Artificial Intelligence in Medicine*, vol. 25, n° 2, 2002, p. 123-148.

[GEI 92] GEIGER D., « An Entropy-based Learning Algorithm of Bayesian Conditional Trees », *Uncertainty in Artificial Intelligence : Proceedings of the Eighth Conference (UAI-1992)*, San Mateo, CA, 1992, Morgan Kaufmann Publishers, p. 92-97.

[GIR 95] GIROSI F., JONES M., POGGIO T., « Regularization Theory and Neural Networks Architectures », *Neural Computation*, vol. 7, n° 2, 1995, p. 219-269.

[HEC 92] HECKERMAN D., NATHWANI B., « An Evaluation of the Diagnostic Accurency of Pathfinder », *Comput Biomed Res*, vol. 25, 1992, p. 56-74.

[HEC 94] HECKERMAN D., GEIGER D., CHICKERING M., « Learning Bayesian networks : The combination of knowledge and statistical data », DE MANTARAS R. L., POOLE D., Eds., *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA, USA, juillet 1994, Morgan Kaufmann Publishers, p. 293–301.

[HEC 98] HECKERMAN D., « A Tutorial on Learning with Bayesian Network », JORDAN M. I., Ed., *Learning in Graphical Models*, Kluwer Academic Publishers, Boston, 1998.

[HEN 89] HENRION M., « Some Practical Issues in Constructing Belief Networks », KANAL L. N., LEVITT T. S., LEMMER J. F., Eds., *Uncertainty in Artificial Intelligence 3*, vol. 8 de *Machine Intelligence and Pattern Recognition*, p. 161–174, North-Holland, Amsterdam, 1989.

[JAA 99] JAAKKOLA T., JORDAN M., « Variational Methods and the QMR-DT Database », *Journal of Articial Intelligence*, vol. 10, 1999, p. 291-322.

[JEN 96] JENSEN F., *Introduction to Bayesian Networks*, Springer Verlag, 1996.

[JOR 98a] JORDAN M. I., *Learning in Graphical Models*, Kluwer Academic Publishers, Dordecht, The Netherlands, 1998.

[JOR 98b] JORDAN M. I., GHAHRAMANI Z., JAAKKOLA T. S., SAUL L., « An Introduction to Variational Methods for Graphical Models », JORDAN M. I., Ed., *Learning in Graphical Models*, Kluwer Academic Publishers, Boston, 1998.

[JOU 00] JOUFFE L., MUNTEANU P., « Smart-Greedy+ : Apprentissage hybride de réseaux bayésiens », *Colloque francophone sur l'apprentissage, CAP, St. Etienne*, juin 2000.

[KAP 00] KAPPEN H., WIEGERINCK W., TER BRAAK E., « Decision support for medical diagnosis », MEIJ J., Ed., *Dealing with the data flood. Mining data, text and multimedia*, The Hague : STT/Bewetong (Study centre for Technology Trends, 65)., 2000.

[KAP 02] KAPPEN H., « The cluster variation method for approximate reasoning in medical diagnosis », NARDULLI G., STRAMAGLIA S., Eds., *Modeling Bio-medical signals*, World-Scientic, 2002.

[KEO 99] KEOGH E., PAZZANI M., « Learning Augmented Bayesian Classifiers : A Comparison of Distribution-based and Classification-based Approaches », *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics*, 1999, p. 225-230.

[KRA 98] KRAUSE P. J., « Learning Probabilistic Networks », 1998.

[LAR 96] LARRANAGA P., KUIJPERS C., MURGA R., YURRAMENDI Y., « Learning Bayesian Network Structures by searching the best order ordering with genetic algorithms », *IEEE Transactions on System, Man and Cybernetics*, vol. 26, 1996, p. 487-493.

[LAU 92] LAURITZEN S., « Propagation of Probabilistics, Means and Variances in Mixed Graphical Association Models », *Journal of the American Statistical Association*, vol. 87, 1992, p. 1098-1108.

[LAV 97] LAVRAC N., KERAVNOU E., ZUPAN B., *Intelligent Data Analysis in Medicine and Pharmacology*, Kluwer, 1997.

[LAV 99] LAVRAC N., « Selected techniques for data mining in medicine », *Artificial Intelligence in Medicine*, vol. 16, n° 1, 1999, p. 3-23.

[LEP 92] LEPAGE E., AL., « Système D'aide à la Décision Fondé sur un Modèle de Réseau Bayésien Application à la Surveillance Transfusionnelle », *Informatique et santé*, vol. 5, 1992, p. 76-87.

[LER 98] LERAY P., Apprentissage et Diagnostic de Systemes Complexes : Réseaux de Neurones et Réseaux Bayesiens. Application à La Gestion En Temps Réel Du Trafic Téléphonique Français, PhD thesis, Université Paris 6, 1998.

[MID 91] MIDDLETON B., SHWE M., HECKERMAN D., HENRION M., HORVITZ E., LEHMANN H., COOPER G., « Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base : Part II. Evaluation of diagnostic performance », *SIAM Journal on Computing*, vol. 30, 1991, p. 256–267.

[MIL 82] MILLER R., POPLE H., MYERS J., « INTERNIST-1, An Experimental Computer-based Diagnostic Consultant for General Internal Medicine », *N Engl J Med*, vol. 307, 1982, p. 468-476.

[MUN 01] MUNTEANU P., BENDOU M., « The EQ Framework for Learning Equivalence Classes of Bayesian Networks », *Proceedings of the First IEEE International Conference on Data Mining, IEEE ICDM*, 2001.

[MUR 01] MURPHY K., « The BayesNet Toolbox for Matlab », *Computing Science and Statistics : Proceedings of Interface*, vol. 33, 2001.

[NEA 98] NEAL R. M., HINTON G. E., « A View of the EM algorithm that justifies incremental, sparse and other variants », JORDAN M. I., Ed., *Learning in Graphical Models*, Kluwer Academic Publishers, Boston, 1998.

[ONI 00] ONISKO A., DRUZDZEL M. J., WASYLUK H., « Learning Bayesian network parameters from small data sets : Application of Noisy-OR gates », *Working Notes of the Workshop on Bayesian and Causal Networks : From Inference to Data Mining, 12th European Conference on Artificial Intelligence (ECAI-2000)*, Berlin, Germany, 2000.

[PEA 86] PEARL J., « Fusion, Propagation, and Structuring in Belief Networks », *Artificial Intelligence*, vol. 29, 1986, p. 241-288.

[PEA 88] PEARL J., *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference.*, Morgan Kaufmann, 1988.

[PEA 91] PEARL J., VERMA T. S., « A Theory of Inferred Causation », ALLEN J. F., FIKES R., SANDEWALL E., Eds., *KR'91 : Principles of Knowledge Representation and Reasoning*, San Mateo, California, 1991, Morgan Kaufmann, p. 441–452.

[PEA 00] PEARL J., *Causality : Models, Reasoning, and Inference*, Cambridge University Press, Cambridge, England, 2000.

[PRA 94] PRADHAN M., PROVAN G., MIDDLETON B., HENRION M., « Knowledge Engineering for Large Belief Networks », *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–94)*, San Francisco, CA, 1994, Morgan Kaufmann Publishers, p. 484–490.

[QUI 86] QUINLAN J., « Induction of decision trees », *Machine Learning*, vol. 1, 1986, p. 81-106.

[REN 01] RENOOIJ S., « Probability Elicitation for Belief Networks : Issues to Consider », *Knowledge Engineering Review*, vol. 16, n° 3, 2001, p. 255-269.

[ROB 77] ROBINSON R. W., « Counting unlabeled acyclic digraphs », LITTLE C. H. C., Ed., *Combinatorial Mathematics V*, vol. 622 de *Lecture Notes in Mathematics*, Berlin, 1977, Springer, p. 28–43.

[SAC 02] SACHA J., GOODENDAY L., CIOS K., « Bayesian learning for cardiac SPECT image interpretation », *Artificial Intelligence in Medecine*, vol. 26, 2002, p. 109-143.

[SCH 78] SCHWARTZ G., « Estimating the dimension of a model », *The Annals of Statistics*, vol. 6, n° 2, 1978, p. 461-464.

[SHO 74] SHORTLIFFE E. H., MYCIN : A Rule-Based Computer Program for Advising Physicians Regarding Antimicrobial Therapy Selection, PhD thesis, Stanford Artificial Intelligence Laboratory, Stanford, CA, octobre 1974.

[SHW 91] SHWE M., MIDDLETON B., HECKERMAN D., HENRION M., HORVITZ E., LEHMANN H., COOPER G., « Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base : Part I. The probabilistic model and inference algorithms », *SIAM Journal on Computing*, vol. 30, 1991, p. 241–250.

[SIE 98] SIERRA B., LARRANAGA P., « Predicting survival in malignant skin melanoma using Bayesian networks automatically induced by genetic algorithms. An empirical comparison between different approaches », *Artificial Intelligence in Medicine*, vol. 14, n° 1-2, 1998, p. 215-230.

[SIE 00] SIERRA B., INZA I., LARRANAGA P., « Medical Bayes Networks », *Lecture Notes in Computer Science*, vol. 1933, 2000, p. 4-14, Springer-Verlag.

[SIE 01] SIERRA B., SERRANO N., LARRANAGA P., PLASENCIA E. J., INZA I., JIMENEZ J. J., REVUELTA P., MORA M. L., « Using Bayesian networks in the construction of a bi-level multi-classifier. A case study using intensive care unit patients data », *Artificial Intelligence in Medicine*, vol. 22, n° 3, 2001, p. 233-248.

[SPI 93] SPIRTES P., GLYMOUR C., SCHEINES R., *Causation, prediction, and search*, Springer-Verlag, 1993.

[SPI 00] SPIRTES P., GLYMOUR C., SCHEINES R., *Causation, Prediction, and Search*, The MIT Press, 2 édition, 2000.

[SRI 93] SRINIVAS S., « A Generalization of the Noisy-Or Model », HECKERMAN D., MAMDANI A., Eds., *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence*, San Mateo, CA, USA, juillet 1993, Morgan Kaufmann Publishers, p. 208–218.

[STE 97] STEIMANN F., « Fuzzy set theory in medicine », *Artificial Intelligence in Medicine*, vol. 11, n° 1, 1997, p. 1-7.

[SUZ 99] SUZUKI J., « Learning Bayesian Belief Networks Based on the MDL Principle : An Efficient Algorithm Using the Branch and Bound Technique », *IEICE Transactions on Information and Systems*, vol. E82-D, n° 2, 1999, p. 356–367.

[SZO 82] SZOLOVITS P., *Artificial Intelligence in Medicine*, Westview Press, Inc., Boulder, Colorado (http ://medg.lcs.mit.edu/ftp/psz/AIM82/), 1982.

[VLA 02] VLASSIS N., LIKAS A., « A greedy EM algorithm for Gaussian mixture learning », *Neural Processing Letters*, vol. 15, 2002, p. 77-87.

[WIE 99] WIEGERINCK W., KAPPEN H., BRAAK E., BURG W., NIJMAN M., NEIJT Y., « Approximate inference for medical diagnosis », *Pattern Recognition Letters*, vol. 20, 1999, p. 1231-1239.

[WU 01] WU X., LUCAS P., KERR S., DIJKHUIZEN R., « Learning Bayesian-Network Topologies in Realistic Medical Domains », *ISMDA*, 2001, p. 302-308.