Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ BADJI MOKHTAR-ANNABA

جامعة باجي مختار – عنابـــــــــة

Faculté des Sciences de l'Ingénieur

Département d'Informatique

Année : 2018/2019

THESE

Présentée en vue de l'obtention du diplôme de

DOCTORAT 3ième cycle

Maintenance de réseau bayésien pour le diagnostic des pathologies des seins.

Filière : Informatique

Spécialité : Traitement d'Image et Vision Artificielle

Par

Ahlem REFAI-BOUSRI

Directeur de thèse :

MEROUANI Hayet Farida — Professeur. Université Badji Mokhtar-Annaba

Devant le jury

Président

GHOUALMI Nacira — Professeur. Université Badji Mokhtar-Annaba

Examinateurs :

SERIDI Hamid — Professeur. Université- 8 Mai 1945-Guelma

MAZOUZI Smaine — MCA. Université- 20 Aout 1955 à Skikda

AZIZI Nabiha — MCA. Université Badji Mokhtar-Annaba

REMERCIEMENTS

Je tiens à exprimer toute ma reconnaissance à mon Directeur de thèse Madame Merouani Hayet Farida Professeur à l'université De Annaba. Je la remercie de m'avoir encadré, orienté, aidé et conseillé.

J'adresse mes sincères remerciements à Monsieur Hamid Séridi, professeur à l'Université 8 Mai 1945 à Guelma ; Mr Smaine Mazouzi, Docteur à l'Université 20 Aout 1955 à Skikda ; Mme Nabiha Azizi Docteur à l'Université de Annaba pour avoir accepté d'être examinateurs de cette thèse.

Je remercie également Mme Nacira Ghoualmi, professeur à l'université de Annaba d'avoir assumé le rôle de présidant du jury.

Je remercie Mme Hayet Aouras Professeur au Service de Maternité CHU à Ibn Rochd ainsi, à la Faculté de Médicine à Annaba, d'avoir donné de son temps pour que je passe faire la collecte des données.

Je remercie mes très chers parents, qui ont toujours été là pour moi, « Vous avez tout sacrifié pour vos enfants n'épargnant ni santé ni efforts. Vous m'avez donné un magnifique modèle de labeur et de persévérance. Je suis redevable d'une éducation dont je suis fière ».

Mes remerciements les plus affectueux s'adressent à mes frères et ma sœur pour leur encouragement.

Je tiens à remercier *Rachida*, pour son amitié, son soutien inconditionnel et ses encouragements,

Enfin merci à tous mes Amies pour leur sincère amitié et confiance, et à qui je dois ma reconnaissance et mon attachement. À tous ces intervenants, je présente mes remerciements, mon respect et ma gratitude.

Merci mon Dieu de m'avoir donné la force, la patience et la volonté d'arriver au terme de travail.

A la mémoire de mon frère Ilyes ; Aucune dédicace ne saurait exprimer l'amour, l'estime, le dévouement et le respect que j'ai toujours eu pour vous. Je ne fais que réalisé votre décès mais vous resterez toujours dans mon cœur.

A mon mari, à mes enfants : Adem et Ines, à toute la famille.

ABSTRACT

This study investigates a proposal of new Bayesian network model for the diagnosis of the most frequent breast pathologies and their implementation under a medical diagnostic system as part of maintenance. It consists in reproducing the process of doctor's diagnosis that allows the identification of a disease through its symptoms.

The proposed Bayesian network allows a representation of qualitative and quantitative knowledge expressing the uncertainty divided into four levels: clinical level, medical imaging level, biological level and diagnostic level. Bayesian networks are used to calculate the probabilities of the most likely posteriori or causes, of an observed anomaly by using the clustering (JLO) algorithm proposed by GeNIe tool and should be sufficient for our application. In order to improve the performances of the system and due to errors in the construction of the model which is supplied a priori by an expert, and the changes in the dynamics domains, we propose the maintenance of BN that implements the policies of updating a fixed structure and considers its reorganization by defining supplementary variables noted as maintenance actions that could be add them or deleted and their values can be edited.

Keywords: Bayesian network, Bayesian inference, Clustering algorithm, Aid to medical decision, Structural update of BN, Learning parameters in GeNIe tool.

RESUME

Cette étude porte sur une proposition d'un nouveau modèle de réseau bayésien pour le diagnostic des pathologies les plus fréquentes du sein et leur implémentation sous un système de diagnostic médical dans le cadre de la maintenance, l'idée principale consiste à reproduire le processus de diagnostic du médecin qui permet une identification d'une maladie par ses symptômes.

Le réseau bayésien proposé permet une représentation des connaissances qualitatives et quantitatives exprimant l'incertitude décomposée en quatre niveaux, niveau clinique, niveau imagerie médicale, niveau biologique et niveau de diagnostic. Nous utilisons les réseaux bayésiens pour calculer les probabilités à postériori de ou des causes, les plus probables d'une anomalie observée en utilisant l'algorithme d'inférence exacte clustering (JLO). En effet, en raison d'erreurs dans la construction du modèle fourni à priori par un expert, et des changements dans la dynamique des domaines, nous avons opté pour une amélioration des performances du système. Nous proposons la maintenance du réseau bayésien qui met en application des politiques de mise à jour d'une structure fixe. La maintenance s'intéresse à sa réorganisation de cette structure en définissant des variables supplémentaires notées comme des actions de maintenance où on pourrait les ajouter ou les supprimer et modifier leurs valeurs.

Mot clé : réseau bayésien, inférence bayésienne, algorithme clustering (JLO), aide à la décision médicale, mise à jour structurelle de RB, apprentissage des paramètres dans l'outil GeNie.

ملخص

تبحث هذه الدراسة في اقتراح نموذج شبكة بايزي جديد لتشخيص أمراض الثدي الأكثر شيوعا وتنفيذها في إطار نظام التشخيص الطبي كجزء من الصيانة. وهو يتألف من استنساخ عملية تشخيص الطبيب الذي يسمح بتحديد المرض من خلال أعراضه.

تسمح الشبكة البايزية المقترحة بتمثيل المعرفة النوعية والكمية التي تعبر عن عدم اليقين مقسمة إلى أربعة مستويات: المستوى السريري، ومستوى التصوير الطبي، والمستوى البيولوجي، ومستوى التشخيص.

وتستخدم شبكات بايزي لحساب الاحتمال الاستدلالي لسبب أو لمجموعة من الأسباب لحالة ملحوظة باستخدام خوارزمية (JLO) المقترحة من قبل أداة GeNie وينبغي أن تكون كافية لتطبيقنا. من أجل تحسين أداء النظام وبسبب الأخطاء في بناء النموذج الذي يتم توفيره مسبقا من قبل خبير، والتغيرات في المجالات الديناميكية، نقترح صيانة الشبكة البايزية التي تطبق سياسات تحديث بنية ثابتة، وتعمل على إعادة تنظيمها من خلال تحديد المتغيرات التكميلية وأشار إلى الإجراءات الصيانة التي يمكن أن تكون إضافتها أو حذفها، ويمكن أن تعدل قيمهم من خلال تحديد متغيرات إضافية المشار إليها بوصفها عقدة إجراء الصيانة التي يمكن إضافتها، حذفها أو تعديل قيمتها.

الكلمة المفتاحية: الشبكة البايزية، الاستدلال البايزي، خوارزمية (JLO)، المساعدة في القرار الطبي، تحديث هيكلية الشبكة البايزية، استدراك العوامل باستخدام GeNie

TABLES DES MATIERES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ACRONYMES

RB	Réseau bayésien
GOA	Graphe orienté acyclique
SADM	Système d'aide à la décision médicale
SADDM	Système d'aide à la décision pour le diagnostic médical
TPC	Tables de probabilités conditionnelles
MV	Maximum de vraisemblance
EM	Expectation Maximisation
JLO	Jensen, Lauritzen, Olessen
GAD	Un graphe acyclique dirigé
MPE	Most probable explanation
SPI	Symbolic Probabilistic Inference
DSL	Decision Systems Laboratory
MFC	Microsoft Foundation Classes
BNJ	Bayesian network tools in Java
KDD	Knowledge Discovery in Databases
GNU	General Public Licence
AIS	Adaptive Importance Sampling
SADMsein	Système Aide au Diagnostic des Maladies des Seins
ODBC	Open Database Connectivity
MAP	Maximum à posteriori
MFC	Microsoft Foundation Classes
SGBD	Système de gestion de base de données
RAS	Rien à signalé
ADP	Adénopathie

Introduction générale

Contexte et problématique

L'un des buts de l'intelligence artificielle (IA) est de concevoir des systèmes capables de reproduire le comportement de l'être humain dans ses activités de raisonnement. L'IA c'est focalisée sur la manière de représenter les connaissances d'un expert et de modéliser son processus de décision, pour construire des systèmes dont le résultat pouvait être qualifié d'« intelligent ». Du fait de l'accroissement continu des connaissances médicales, les signes, les symptômes et les maladies se sont spécialisés, les investigations complémentaires se sont multipliées, de nombreuses nouvelles molécules sont arrivées sur le marché. Les médecins, qu'ils soient généralistes ou spécialistes, ne peuvent plus maîtriser l'ensemble du savoir médical permettant de reconnaître les maladies ou de déterminer la meilleure prise en charge thérapeutique.

Aussi, ils ont souvent recours à des sources d'information externes, traditionnellement les collègues, les livres, et la littérature scientifique, pour trouver les informations manquantes. Néanmoins, en dépit (ou à cause ?) de la diffusion en ligne de grands volumes de ressources documentaires facilement accessibles, la recherche de la solution au problème posé par un patient donné est une tâche difficile. Très tôt, les qualités de l'ordinateur (mémoire, rapidité, puissance de calcul) se sont imposées comme des solutions potentielles à cette difficulté et des systèmes informatiques d'aide à la décision médicale ont été développé allant de la gestion des établissements de soins et des cabinets médicaux, à la mise au point des systèmes experts, en passant par les systèmes d'aide au diagnostic.

Parmi les différents types de systèmes d'aide à la décision médicale (SADM), on distingue, les systèmes qui s'appuient sur des modélisations mathématiques permettant de produire des probabilités à partir d'un jeu de données (probabilité d'un diagnostic, de la survenue d'un événement grave, etc.). Historiquement, les premiers SADM étaient des systèmes d'aide au diagnostic. Ils utilisaient des approches numériques, essentiellement statistiques et probabilistes. Dans ce travail nous nous s'intéressons aux systèmes d'aide au diagnostic médicale qui utilisent des réseaux bayésien probabilistes dans la représentation des connaissances médicale.

Les premiers programmes informatiques d'aide au diagnostic ont vu le jour au début des années 60 aux Etats-Unis, le professeur Warner proposa dès 1961, un système informatique capable de diagnostiquer 33 maladies cardiaques congénitales, d'après la présence ou l'absence de 50 symptômes et ce à partir de signes constatés sur le patient, le système calculait les probabilités de 33 diagnostics.

En Grande Bretagne, le professeur de Dombal, a mis sur ordinateur, à la fin des années 60, le diagnostic des urgences abdominales (appendicite, occlusion intestinale, etc.…), huit maladies au total, identifiées d'après 50 symptômes [BRA 16].

En peu d'années, les systèmes d'aide au diagnostic se multiplient : diagnostic des maladies pulmonaires, neurologiques, etc...

En générale ces systèmes ont pour but de reproduire l'expérience et le mode de raisonnement des spécialistes. Dans un domaine d'application précis et limité, ils peuvent être utilisés à résoudre des problèmes, vérifier des résultats, satisfaire des requêtes, détecter des incohérences, l'amélioration de la qualité des soins et de la prise en charge des malades par la réduction des erreurs médicales [SER 13].

Motivations et objectifs

Le cancer du sein est la deuxième cause de décès et la maladie la plus fréquentes chez les femmes à nos jours. A ce titre plusieurs travaux ont été effectués afin de développer des outils d'aide au diagnostic de cette maladie cancéreuse.

Le diagnostic est une démarche incontournable pour prendre une décision et mettre en route un traitement. Cette décision est souvent prise à partir d'informations incertaines et/ou incomplètes, elles sont difficiles à synthétiser [CLE 01] et donc il est fréquent de tirer des conclusions incertaines ce qui explique le développement des méthodes probabilistes. Nous nous appuyons sur les réseaux bayésiens qui ont marqué leurs présences dans l'aide au diagnostic médicale.

Depuis quelques années, les recherches utilisant les réseaux bayésiens tournent autour des axes principaux, et qui sont abordés aussi dans le cadre de notre étude :

- La construction du réseau bayésien.
- Les algorithmes de propagation d'informations dans le réseau, autrement dit l'inférence probabiliste.
- L'utilisation pratique des réseaux bayésiens dans des problèmes médicaux.

- Les outils de manipulation graphique des réseaux bayésien.

Des tentatives de mise en place de systèmes d'aide au diagnostic à base de réseau bayésien ont été faites, par une proposition d'un nouveau modèle construit avec l'outil GeNIe qui permet la saisie, la modification, l'utilisation et l'apprentissage de modèles à base de réseau bayésien.

Les systèmes d'aide au diagnostic nécessitent en phase de fonctionnement d'être maintenu ; en raison d'erreurs dans la construction du modèle qui est fournie a priori par un expert et des changements dans la dynamique des domaines afin de garantir la qualité de ce système. Nous proposons, alors, la maintenance du RB qui met en application des politiques de mise à jour d'une structure fixe et s'intéresse à sa réorganisation en définissant des variables supplémentaires notées comme des actions de maintenance. Cependant, la mise à jour du réseau est encore un problème ouvert nous étudions la maintenance de la structure et les données des paramètres qui sont les deux susceptibles de changer.

Les objectifs de notre travail se résument aux points suivants :

- Proposer un nouveau modèle de Réseau bayésien pour l'aide au diagnostic des maladies des seins.
- Collecter des dossiers médicaux par l'étude statistique à l'hôpital Ibn Roched d'ANNABA.
- Développer un système d'aide au diagnostic des pathologies les plus fréquentes des seins.
- Obtenir un diagnostic le plus précis possible et avec la plus grande certitude en implémentant le modèle sous un outil de réseau bayésien qui permet la représentation des données et de la connaissance médicale.
- Améliorer les performances du système de diagnostic en développant une nouvelle stratégie de maintenance bayésienne qui consiste à définir des variables supplémentaires notées comme des actions de maintenance pour mettre à jour le réseau bayésien en raison d'erreurs dans le modèle construit et des changements dans la dynamique du domaine.

Organisation du document

Les travaux réalisés dans cette thèse, se composent de cinq chapitres :

Dans le premier chapitre nous introduisons les principes fondamentaux des systèmes d'aide au diagnostic médical, nous donnons les notions de base concernant le diagnostic médical, les différents types des SADM et leurs champs d'application ainsi que les principaux systèmes existants.

Le deuxième chapitre présente une synthèse sur les réseaux bayésiens. Il présente les principales définitions et les principaux théorèmes et les étapes de construction des réseaux bayésiens. Ce chapitre donne une idée globale sur les réseaux bayésiens en abordant les points nécessaires comme l'inférence afin de mieux cerner ce domaine.

Dans le troisième chapitre, nous décrirons certains des principaux logiciels conçus pour les applications de réseau bayésien, les domaines d'application utilisés ainsi que le contexte de l'incertitude médicale.

Dans le quatrième chapitre, nous allons aborder l'aspect conception du système proposé SADMsein (Système Aide au Diagnostic des maladies des seins). Nous sommes concentrés premièrement sur la réalisation du modèle bayésien avec l'assistance du médecin au moyen d'une discussion qui nous a menées à déterminer les variables représentant les maladies du sein leurs principaux signes, symptômes, les tests, et les examens. Pour améliorer les performances du système de diagnostic nous avons rajouté au réseau bayesien une partie additionnelle qui consiste à développer une nouvelle stratégie de maintenance bayésienne en définissant des variables supplémentaires notées comme des actions de maintenance. Avec cette maintenance nous pourrions mettre à jour le réseau bayésien, car des erreurs s'introduiront que ce soit lors de sa construction ou lors des changements apportés par la dynamique du domaine.

Le dernier chapitre est consacré à la phase de réalisation de notre outil support, nous présentons une application pour l'aide au diagnostic des pathologies des seins avec le logiciel GeNie, suivi de la présentation et de la discussion des résultats. Cette application destinée principalement aux médecins ainsi qu'aux praticiens pour faciliter le diagnostic et la prise en charge du cancer du sein, et pour aider également les étudiants en médecine dans l'enseignement en leur offrant cette application comme complément pédagogique.

Ce travail se termine par une conclusion générale et des perspectives.

Chapitre 1 Aide à la décision pour le diagnostic médical

1.1. Introduction

Dans son sens courant, le terme décision renvoie au choix d'une action à faire. Etant donné la grande complexité des problèmes de décision, il est souvent utile de faire appel à une aide extérieure pour la prise de décision. L'aide à la décision n'a pas pour but de remplacer le décideur en lui proposant des solutions « toutes faites ». Elle cherche plutôt à le guider vers des décisions qu'il aura à prendre sous sa responsabilité. L'aide à la décision se rencontre dans de nombreux domaines applicatifs tels que l'économie, les mathématiques, l'informatique, la médecine, etc.

En médecine, la décision est considérée comme étant le centre de l'acte médical. Le processus de la décision médicale consiste entre autres à poser **un diagnostic**, **proposer un traitement ou le différer, etc.**

Le diagnostic médical est une action de décision qui suppose sur le recueil des données à travers les symptômes, les résultats d'examens biologiques ou radiologiques, un signal, une séquence vidéo [MAL 16], ce sont les connaissances du médecin qui permettent de valider un diagnostic fiable. Les systèmes d'aide à la décision pour le diagnostic médical ou (*système d'aide au diagnostic tout court*) nécessitent en particulier une représentation adéquate des connaissances mises en jeu, ainsi que des mécanismes efficaces d'exploitation de ces connaissances, ou de raisonnement.

Le raisonnement en intelligence artificielle concerne l'ensemble des techniques permettant la manipulation des connaissances, déjà acquises afin de produire de nouvelles connaissances Pour un système intelligent, le raisonnement est en général conditionné par le but que l'on souhaite atteindre pour résoudre un problème donné. Ainsi, ce type de raisonnement ne déduit pas l'ensemble des connaissances mais seulement la partie des connaissances intéressantes qui sont associées au but recherché. Différents mécanismes de raisonnement sont utilisés en intelligence artificielle pour produire de nouvelles connaissances. Ainsi, le raisonnement peut être qualifié en fonction de sa nature, raisonnement par déductions statistiques, raisonnement à base de cas, par contraintes, etc.) ; Bien qu'il existe différents modes de représentation des connaissances possibles (représentations logiques, réseaux sémantiques, règles de production, réseau bayésien etc.) mais le choix du mode de représentation prend en compte la diversité des connaissances mises en œuvre. Les données médicales souffrent, en général, au moins d'un type d'imperfection comme par exemple l'imprécision, l'incertitude [ALS 12].

Notons que la majorité des systèmes d'aide au diagnostic médical se contente d'utiliser un seul des différents modes de représentation des connaissances et du raisonnement associé. Notre travail s'inscrit dans le domaine des systèmes d'aide à la décision pour le diagnostic médical (SADDM) dans l'objectif de réaliser une exploitation d'un raisonnement par déductions statistiques, afin de mettre à la disposition du médecin un ensemble d'informations destiné à l'aider au cours de son processus de prise de décision diagnostique.

Dans ce chapitre, nous présentons les principes fondamentaux des systèmes d'aide à la décision médicale, ainsi les différentes topologies d'un SADM, nous accentuons sur les systèmes d'aide à la décision pour le **diagnostic** médical en donnant les notions de base concernant le diagnostic médical ainsi que les principaux systèmes existants.

1.2. Notion de décision médicale

En médecine, la décision est considérée comme étant le centre de l'acte médical. Le processus de la décision médicale consiste entre autres à poser un diagnostic, proposer un traitement ou le différer, etc. Ainsi, de très nombreuses applications d'aide à la décision ont été développées dans ce domaine. Ces applications sont destinées à soutenir le personnel de santé dans leurs prises de décision. Cela implique l'utilisation de divers outils d'aide à la décision [DAR 03].

1.3. Qu'est-ce qu'un système d'aide à la décision médicale (SADM)?

Plusieurs définitions concernant les systèmes d'aide à la décision médicale ont été proposées :

- Un système d'aide à la décision médicale est un ensemble organisé d'informations, conçu pour assister le praticien dans son raisonnement en vue d'identifier un diagnostic et de choisir la thérapeutique adéquate, en opérant un dialogue entre l'homme et la machine [DAR 03].
- Les systèmes d'aide à la décision médicale (SADM) sont « des applications informatiques dont le but est de fournir aux cliniciens en temps et lieux utiles les informations décrivant la situation clinique d'un patient ainsi que les connaissances appropriées à cette situation, correctement filtrées et présentées afin d'améliorer la qualité des soins et la santé des patients » [LOB 07].

1.4. Les champs d'application des SADM

Les champs d'applications des SADM sont multiple. ils concernent la plupart des domaines médicaux [GAU 14].

A. L'ensemble des activités médicales :
   - La prévention
   - Le dépistage
   - Le diagnostic
   - La prescription

B. La plupart des spécialités médicales
   - Le suivi des maladies chroniques
   - Les affectations aigues
   - Les urgences

C. L'ensemble des médecins
   - Médecins généralistes
   - Médecins spécialistes
   - Médecins en formation

D. Les différents modes d'exercice
   - Cabinets
   - Hôpital
   - Public ou privé.

1.5. Les composants d'un système d'aide à la décision médicale

Les différents types de SADM se composent de :

- D'une base de connaissances,
- D'un moteur d'inférence ou d'exécution,
- D'interfaces assurant la communication avec l'utilisateur, le dossier médical et le logiciel métier.

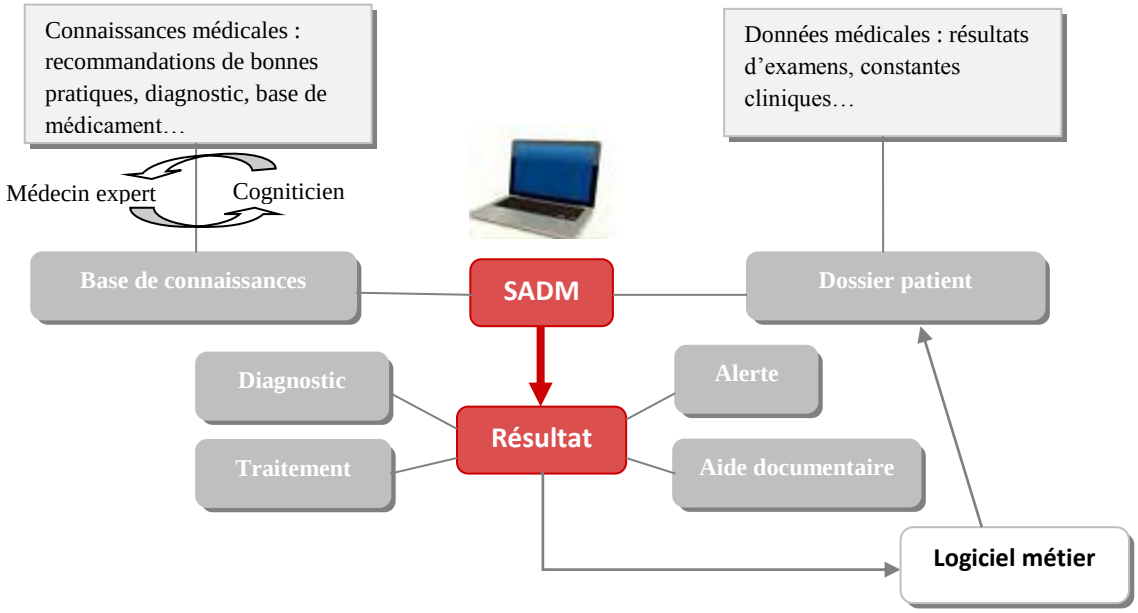


Fig.1. 1 Schéma d'un SADM

1.6. La typologie des systèmes d'aide à la décision médicale

Il existe trois grandes catégories de systèmes d'aide à la décision médicale.

1.6.1. Les systèmes d'assistance documentaire

L'objectif des systèmes d'assistance documentaire est de faciliter l'accès aux informations pertinentes en un temps record mais ces systèmes n'ont pas de méthode de raisonnement à proprement parler [CLE 01].

L'accès aux résultats de laboratoire, la consultation des éléments importants du dossier médical du patient, les références bibliographiques (par exemple Medline) et les systèmes de bases de données concernant les médicaments constituent des aides indirectes à la décision. Cette aide intervient pour faciliter l'appréciation d'une situation par le médecin [DEG 91].

1.6.2. Les systèmes d'alerte ou de rappels automatiques

Certains systèmes permettent de rappeler au médecin des erreurs à ne pas commettre ou des éléments importants à prendre en compte pour la décision. Ils sont plus actifs et plus directement impliqués dans la décision médicale. L'assistance fournie n'est pas une aide au raisonnement ou à l'appréhension globale du cas du patient, mais plutôt un aide-mémoire fournissant une information utile et pertinente dans une situation facile à définir a priori (Degoulet& Fieschi, 1991). Ces systèmes, comme les précédents, ne raisonnent pas véritablement [CLE 01].

Par exemple, le rappel des valeurs normales des résultats d'examens de biologie et l'utilisation d'une typographie permettant d'attirer l'attention sur les valeurs anormales ou encore l'émission d'un message de mise en garde devant une association de médicaments déconseillée constituent une aide simple dont l'utilité est primordiale [DEG 91].

1.6.3. Les systèmes consultants

Face à une situation médicale bien définie telle que : **un diagnostic**, **une thérapie** ou **un pronostic**, les systèmes consultants tentent d'émettre un avis de spécialiste. Ces systèmes fournissent à [DEG 91] l'utilisateur des conclusions argumentées selon les méthodes de raisonnement employées. Dans cette catégorie on s'intéresse principalement aux ***systèmes d'aide à la décision pour le diagnostic médical***.

1.7. Diagnostic médical

Le mot « diagnostic » provient du grec διάγνωση, diágnosi, à partir de δια-, dia-, par, à travers, séparation, distinction et γνώση, gnósi, la connaissance, le discernement ; il s'agit donc d'acquérir la connaissance à travers les signes observables. Cette définition introduit naturellement la notion de catégories ou classes diagnostiques préexistantes, l'instance à classer et le jugement que l'instance appartient à une classe plutôt qu'à une autre [KIN 67]. Le diagnostic médical a été défini par Jean-Charles Sournia dans [SOU 95] comme suit :

« Démarche intellectuelle par laquelle une personne d'une profession médicale identifie la maladie d'une autre personne soumise à son examen, à partir des symptômes et des signes que cette dernière présente, et à l'aide d'éventuelles investigations complémentaires ».

En effet, un diagnostic médical représente une tâche difficile à réaliser parce qu'il repose sur la capacité de raisonnement du médecin et de son aptitude à prendre des décisions alors que les informations utilisées sont potentiellement entachées d'incertitude et d'autres formes d'imperfection. L'incertitude est d'origine multiple : possibilité d'erreur dans les données, ambigüité de la représentation de l'information, incertitude sur les relations entre les diverses informations [LEP 92]. Cette difficulté a conduit à la conception et au développement de systèmes d'aide au diagnostic ayant pour but d'assister les médecins dans l'élaboration de leurs diagnostics.

1.7.1. La notion de diagnostic médical

La médecine n'est pas seulement une discipline scientifique mais elle est également une discipline d'action qui requiert souvent une prise de décision. Ce processus résulte de la

confrontation d'un problème réel à l'expérience acquise et à un corpus de connaissances théoriques.

Un diagnostic médical représente l'acte d'associer le nom d'une ou plusieurs maladies ou syndromes à des manifestations observées (antécédents, symptômes, signes) dans un cas de patient [MIL 09]. Le processus de diagnostic médical se déroule comme suit :

Premièrement, le médecin constate les symptômes se manifestant chez un patient. A partir de ces symptômes, il formule des hypothèses diagnostiques initiales. Dans un deuxième temps, il procède à un examen initial du patient, qui lui permet d'augmenter la part de confiance pour certaines hypothèses, et la diminuer pour d'autres. En même temps, le médecin pose au patient des questions dont les réponses peuvent être utiles à conforter ou rejeter une hypothèse initialement formulée.

Le médecin « réalise » une mise en correspondance entre les informations obtenues au cours des trois étapes précédentes avec les connaissances qu'il possède de part sa formation et son expérience. Si, au terme des étapes précédentes, le taux de confiance d'une certaine hypothèse s'accroît au point de dissiper le doute sur la maladie à laquelle est confronté le médecin, ce dernier peut alors formuler son diagnostic final et prescrire le traitement adéquat au patient. Si le cas reste ambigu après les trois étapes indiquées, le médecin cherche alors une autre source d'informations qui puisse apporter une quantité d'informations supplémentaires permettant d'éliminer l'ambiguïté [HOM 12]. Souvent, il demande une analyse complémentaire qui peut être sous forme d'analyses sanguines, d'imagerie médicale, etc. Il acquiert de l'information supplémentaire qui vient compléter la quantité d'informations dont il dispose déjà, et qui lui permettent de confirmer ou d'infirmer la ou les hypothèses qu'il a déjà énoncées. Si le médecin n'arrive toujours pas à établir un diagnostic fiable, une dernière étape consiste à ce qu'il ait recours à l'étude d'une base de cas similaires traitées par le passé afin d'établir une correspondance avec le cas actuel auquel il est confronté en s'appuyant sur toutes les informations dont il dispose. Il utilise alors les cas les plus similaires (leurs solutions) afin d'en extraire des informations l'aidant à trouver une solution à son cas.

En effet, il est très clair que le processus de diagnostic médical repose sur la capacité de raisonnement du médecin et de son aptitude à prendre des décisions alors que les informations utilisées sont généralement hétérogènes (examen clinique, images, tests de laboratoire, signaux, vidéos, etc.), et potentiellement entachées d'incertitudes. Ces incertitudes sont d'origines multiples : les informations utilisées peuvent être ambiguës car le malade peut

exprimer une complainte et le médecin en entendre une autre. Ces informations peuvent être incomplètes car, en situation de prise de décisions, le médecin doit agir sans connaitre l'ensemble des données relatives à un patient et bien entendu toute la connaissance spécifique de la situation. Elles peuvent être incertaines car les connaissances cliniques peuvent concerner des maladies plus ou moins fréquentes, ayant des formes cliniques différentes et n'exprimant pas toujours la même symptomatologie, partageant certains signes avec d'autres maladies ou présentant des réponses variables à un traitement donné. Ces différentes raisons prouvent que le diagnostic médical est un processus difficile à réaliser, et le médecin a souvent besoin d'aide afin d'établir une décision de qualité. Ce besoin a conduit à la conception et au développement de systèmes d'aide au diagnostic ayant pour but d'assister les praticiens dans l'élaboration de leur diagnostic [MOK 16].

1.7.2. Système d'aide au diagnostic médical

Définition

Plusieurs définitions, du système d'aide au diagnostic, ont été proposées dans la littéraire. Sim et al. [KON 08] [ALE 10] ont proposé la définition suivante :

« Logiciel conçu pour être une aide directe à la prise de décision clinique, Dans lequel les caractéristiques d'un patient individuel correspondent à une base de connaissances cliniques informatisées, et les évaluations ou recommandations spécifiques au patient sont ensuite présentées au clinicien ou au patient pour une décision ». [KAW 05] ont défini le système d'aide au diagnostic comme suit : « Nous avons défini un système de soutien à la décision clinique comme tout système électronique ou non électronique conçu pour aider directement dans la prise de décision clinique, dans lequel les caractéristiques des patients individuels sont utilisées pour générer des évaluations ou des recommandations spécifiques au patient qui sont ensuite présentées aux cliniciens pour examen ».

Ces différentes définitions confirment le fait que l'aide (informations obtenues par le système) fournie au médecin dans son processus de diagnostic, peut prendre plusieurs formes (cas similaires déjà diagnostiqués, diagnostics potentiels, etc.).

En général, les systèmes d'aide au diagnostic médical (SADM) représentent un moyen potentiel pour améliorer la qualité, la sécurité et l'efficience des soins lorsqu'ils sont accessibles aux médecins pendant leurs activités de soins au moyen de leurs outils métier (dossier médical et prescription informatisée) et s'ils sont correctement intégrés au processus de travail clinique. SADM sont utilisés pour les patients hospitalisés ou pour les soins

ambulatoires. Ils sont régulièrement évalués et améliorés en fonction du retour des utilisateurs, et sont aujourd'hui diffusés dans des établissements ou structures de soins ambulatoires des réseaux de santé, dont font partie les Institutions pionnières. Sous l'influence des utilisateurs, les modalités d'intervention tendent à s'éloigner du modèle prescriptif des SADM historiques et évoluent vers un soutien aussi discret que possible du processus cognitif des cliniciens, guidé par le principe « faire en sorte que la décision appropriée soit la plus facile à prendre » [JEA 01].

Les systèmes experts, les systèmes d'apprentissage, les systèmes de fouille de données, les systèmes d'indexation et de recherche d'images, les systèmes de raisonnement à base de cas et les systèmes de raisonnement par classification sont tous des exemples des systèmes d'aide au diagnostic.

Parmi ces différents types de systèmes d'aide au diagnostic proposés dans la littérature, nous nous intéressons particulièrement aux systèmes fondés sur le raisonnement par déduction. Par la suite, nous présentons les principaux systèmes de diagnostic médical.

1.7.3. Objectifs des systèmes d'aide à la décisionpour le diagnostic médical

Il est intéressant de décrire les SADDM selon le besoin ou le problème précis auxquelsils répondent. Il faut considérer l'objectif primaire pour lequel le SADDM [MOR 15] est conçu :

- ✓ Améliorer l'efficacité globale,
- ✓ Dépister une maladie précocement,
- ✓ Eviter des événements indésirables,
- ✓ Préciser un diagnostic complexe ou un traitement en l'inscrivant dans un protocole.
- ✓ Améliorer l'efficience des soins.
- ✓ Minimiser les risques et les couts.

L'élaboration d'un scénario pour réaliser cet objectif dépend de la définition du périmètre d'action du SADDM que l'on confronte à l'intention du clinicien.

1.8. Quelques systèmes d'aide à la décision pour le diagnostic médical

Les systèmes d'aide au diagnostic ont fait leur apparition dans certains domaines, particulièrement en médecine et ce dans plusieurs spécialités. Même si de nombreux outils existent déjà, leur déploiement en pratique reste modeste. Nous abordons quelques systèmes qui sont fondis sur les réseaux bayésiens.

A. Le Leeds abdominal pain system

Le Leeds abdominal pain system est un système d'aide au diagnostic des patients souffrant de douleurs abdominales aigus lancé en 1972 s'appuyant sur un modèle probabiliste [SID 13].

B. Internist

Internist lancé en 1974 est un des systèmes à vocation large : la médecine interne, avec 600 maladies et 4500 signes ; chaque maladie est décrite par des signes dotés d'un coefficient de sensibilité et de spécificité. Les performances de ce système ont été évaluées sur des cas cliniques du New England Journal of Medicine. Cependant, il est inutilisable en pratique en raison notamment du temps de consultation ; une version simplifiée et à vocation didactique, est maintenant disponible sur micro-ordinateur [KUM O5].

C. Pathfinder

Au niveau des pathologies chirurgicales, un seul système d'aide au diagnostic a été pleinement développé : le système Intellipath. Celui-ci est la version commerciale du système Pathfinder, réalisé par Nathwani et Heckerman à la fin des années 80 et début des années 90. Pathfinder est basé sur un réseau bayésien. Il a été initialement développé dans le domaine de la maladie du nœud de la lymphe (aide au diagnostic des pathologies ganglionnaires), un domaine dans lequel il est difficile de faire un diagnostic. Dans ce système, c'est le docteur Nathwani qui détenait l'expertise et DavidHeckermann [HEC 92] détenait la technique à mettre en œuvre. On peut dire que Pathfinder est l'un des systèmes pionnier de la mise en pratique des réseaux bayésiens.

D. DXPlain

DXPlain [HOF 05] est un système d'aide à la décision pour le diagnostic qui intègre près de 2000 pathologies. Il fonctionne selon des principes de logique probabiliste bayésienne et se montre capable à partir d'un ensemble de signes et symptômes qui caractérisent le patient de générer une liste ordonnéede pathologies possibles. (Fig.1.2) est une copie d'écran de la première version de DXPlain. Lancé en 1986 par une équipe du Massachusetts General Hospital, il n'a depuis cessé d'évoluer. Aujourd'hui encore, une version web 1 du système est disponible, principalement pour la formation des étudiants en médecine.

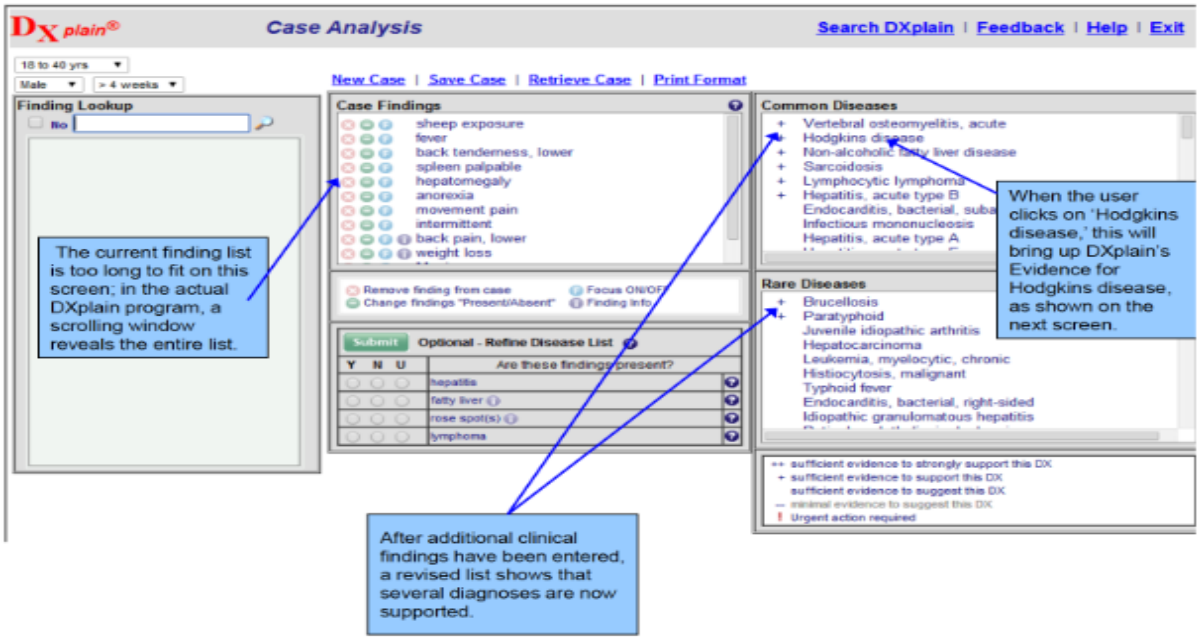


Fig.1. 2Interface de la démo DXplain accessible à http://dxplain.org/demo2/

DXplain propose une liste de diagnostics possibles, en distinguant les diagnostics fréquents et les diagnostics rares. Il peut également poser des questions sur des critères cliniques permettant de discriminer de façon efficace entre les hypothèses diagnostiques retenues.

E. ONCODOC

En 1999 a aussi vu le jour ONCODOC [SER 01], un Système d'aide au diagnostic médical qui repose sur les guides de bonne pratique pour le traitement du cancer du sein. La base de connaissance y est représentée sous la forme d'un "arbre de décision au sein duquel le médecin navigue conformément à sa perception clinique, donc informelle, de critères patient" [SER 04]. La version de 1999 comptait 64 paramètres décisionnels pour un arbre total de 2314 feuilles [SER 99]. Ce système présenté dans la (Fig.1.3) a bien été reçu par le corps médical et a évolué au cours du temps en incluant d'autres types de cancer du sein ainsi que les nouveaux guides de bonne pratique traitant de sujets tels que la chimiothérapie ou l'hormonothérapie.

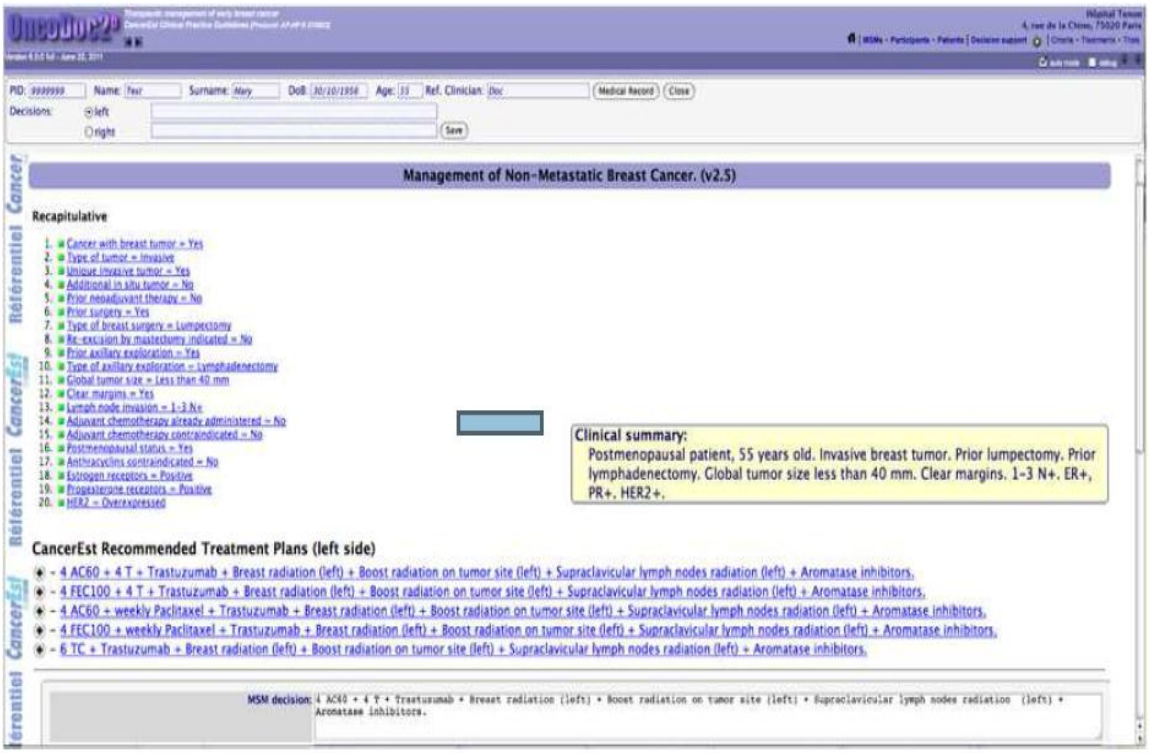


Fig.1. 3Copie d'écran du système OncoDoc2

F. ADM (aide au diagnostic médical)

Il a été développé en France par P.Lenoire [LEN 80]. ADM est similaire à internist dans ses objectifs mais avec un spectre encore plus large que celui de la médecine interne. La base de connaissances contenait 12 000 descriptions de maladies décrites par 130 000 entités et couvrait toute la médecine ainsi que les effets indésirables des médicaments. Chaque maladie, syndrome ou forme clinique, était décrite par un ensemble de signes avec des notions de fréquence et de puissance évocatrice. Le système intégrait également les données de littérature et les dires des experts. Du fait de la largeur de sa couverture, le système n'a pas pu être évalué de manière aussi rigoureuse qu'interniste. Même si sa mise à jour s'est arrêtée au milieu des années 1990, il reste intéressant pour sa partie clinique qui, dans certains domaines, est toujours valide. Il est disponible sous une forme web sur le site du laboratoire d'informatique médicale de l'université de Rennes.

G. RecosDoc-Diabète

Un système d'aide à la décision pour la prise en charge du diabète de type 2, le système permet de construire de façon interactive un profil centré-patient et d'obtenir les recommandations appropriées. Le RecosDoc-diabète [BOU 13] a été publié et mis en ligne en février 2013 comme le montre la (Fig.1.4)

Fig.1. 4 Copie d'écran du système RecosDoc-Diabète

1.9. Lesdifférences mécanisme de raisonnement

Différents mécanismes de raisonnement sont proposés dans la littérature en fonction de laméthodologie utilisée pour la représentation de connaissances. Nous citons :

- Le raisonnement à base de règles [YUA 17] [PEI 13],
- Le raisonnement bayésien,
- Le raisonnement heuristique,
- Les réseaux sémantiques,
- Les réseaux neurones [HOP 88] [MCC88],
- Le raisonnement à base de cas [SCH 82] [SCH 86], etc.

L'efficacité du mécanisme de raisonnement vient généralement de sa capacité de raisonner à partir des différents types d'informations contenant différents types d'imperfections [KON 08].

Les problèmes essentiels à résoudre, afin d'élaborer un système d'aide au diagnostic, sont : la représentation de connaissances et le mode de raisonnement.

En effet, quand nous parlons de l'information (comme terme général) dans le domaine médical, nous ne pouvons jamais négliger les concepts principaux concernant la nature de l'information quantitative (numérique, binaire) et qualitative (nominale ou ordinale), ainsi que l'imperfection de ces informations comprenant l'incertitude, l'imprécision, l'ambigüité, etc. Ces concepts vont être définis le mode de raisonnement utilisé.

Notons que l'information médicale, en général peut être entachée par au moins un type d'imperfection plus précisément l'incertitude ce qui nous a mené à la théorie des probabilités qui est utilisée, depuis longtemps, comme cadre de représentation d'information incertaine, ce qui explique d'une part le développement des méthodes probabilistes comme les réseaux bayésiens introduits par Pearl [PEA 98] [CHE 00]et qui représentent des techniques les plus intéressantes [RAI 07] [SAH 07] [KON 93] de l'intelligence artificielle dans le cadre d'un diagnostic médical ils permettent la représentation des connaissances par un graphe causal intuitif et compréhensible de plus, comme ils sont basés sur des probabilités [BAY 91], ils intègrent l'incertitude dans le raisonnement.

On s'intéresse dans les prochains paragraphes aux réseaux bayésiens utilisées pour le diagnostic médical.

1.10. Nos choix et notre démarche

Nos choix dépendent du domaine choisis et la nature de ses connaissances qui sont souvent imprécises, ambiguës et incomplètes, ainsi leur interprétation et leur utilisation sont de fait incertain ; c'est le cœur du problème de décision médicale ce qui explique d'une part le développement des méthodes probabilistes.

Les réseaux bayésiens représentent le cadre général adopté pour le système d'aide au diagnostic proposé dans ce travail. Ils constituent aujourd'hui l'un des formalismes les plus complets et les plus cohérents pour l'acquisition, la représentation et l'utilisation de connaissances par des ordinateurs, ils décrivent hiérarchiquement toutes les étapes du raisonnement du médecin.

Nous nous intéressons dans notre démarche à la modélisation des connaissances pertinentes qui englobe le problème à étudier pour construire un nouveau modèle de réseau bayésien pour l'aide au diagnostic des pathologies les plus fréquents des seins ainsi la représentation de l'ensemble des données des paramètres du réseau et leurs interprétations sous forme des probabilités. De nombreux outils de manipulation graphique de réseaux

bayésien sont disponibles. Ils sont pour but de permettre la saisie, la modification, l'utilisation et l'apprentissage de modèles à base de réseau bayésien.

1.11. Conclusion

Les systèmes d'aide au diagnostic (SADM) sont des outils informatiques capables de traiter l'ensemble des caractéristiques d'un patient donné afin de générer les diagnostics probables de son état clinique ou les traitements qui lui seraient adaptés (aide à la thérapeutique). Dans ce chapitre, nous avons présenté le contexte général et la problématique de ce travail. En particulier, nous avons présenté la notion de diagnostic médical et la démarche utilisée par le médecin pour établir un diagnostic face à un cas patient donné. Nous avons concentré notre effort sur un type de systèmes d'aide au diagnostic : le système de raisonnement par déductions statistiques. Nous avons présenté quelques exemples des systèmes tels que les systèmes DXPlain, Pathfinder, ONCODOC ect.

Dans le système proposé, la représentation des connaissances et le raisonnement sont définis dans un environnement probabiliste. Le choix de la méthode (réseau bayésien), comme cadre général, a été motivé dans le chapitre suivant.

Chapitre 2 Les réseaux bayésiens

2.1. Introduction

Les réseaux bayésiens, qui doivent leur nom aux travaux de Thomas Bayes au dix-huitième siècle sur « la probabilité des causes », sont le résultat de recherches effectuées dans les années 80.

Les réseaux bayésiens constituent une technique d'acquisition de représentation et de manipulation de connaissance et on les utilise, surtout, pour leur capacité d'effectuer des inférences dans un contexte d'incertitude et aussi pour leurs algorithmes d'apprentissages. Ils sont utilisés pour prévoir, contrôler et simuler le comportement d'un système, à analyser des données et à prendre des décisions, à diagnostiquer les causes d'un phénomène observé grâce à la jonction de la théorie des probabilités et de la théorie des graphes, permettant effectivement de décrire les relations régissant un ensemble de variables aléatoires et d'effectuer un raisonnement probabiliste sur celles-ci. D'ailleurs, les domaines d'applications sont variés. Ils ont été développés pour la première fois en médecine dans le cadre de la prévention, le diagnostic [LER02] [BEL02], et autre, dans l'industrie, la classification automatique de documents structurés ou encore l'analyse d'images et bien d'autres domaines [LAB03].

2.2. Définition de réseau bayésien

Les réseaux bayésiens [JEN96] [PEA88] [NAï07], également appelés réseaux probabilistes, sont des outils de modélisation de connaissances incertaines et complexes. Ils permettent aussi la représentation des relations d'influences entre ces connaissances. Ils ont été utilisés dans de nombreuses applications telles que dans le diagnostic médical [LER04], en bioinformatique [WIL07], corrélation d'alertes [QIN05] [BEN08a], reconnaissance de la parole [ZHO05], filtrage de spams [SAH98], détection d'intrusions [BEN2008b] [KRU03] [BEN04], etc.

Ces modèles de raisonnement probabiliste se caractérisent par deux aspects : un aspect graphique ou qualitatif permettant de représenter d'une manière très simple la connaissance sous forme d'un graphe orienté sans cycles, et un aspect probabiliste ou quantitatif offrant un moyen de quantifier l'incertitude des relations d'influences entre les variables du domaine étudié cet aspect quantitatif est constitué de tables de probabilités dans le cas discret ou de distribution gaussienne dans le cas continu. Plus précisément, un réseau bayésien est défini

comme un graphe orienté acyclique (GOA) permettant de représenter les dépendances directes ou conditionnelles entre les variables du domaine étudié.

Il est muni d'un ensemble de tables de probabilités conditionnelles (TPC) pour quantifier l'incertitude relative aux relations d'influences.

2.3. Définition Formelle

Un réseau bayésien peut être formellement défini par :

• un graphe acyclique orienté G, G=G (V, E) où V est l'ensemble des nœuds de G, et E l'ensemble des arcs de G.

• un espace probabilisé fini (W, p).

• un ensemble de variables aléatoires associées aux nœuds du graphe et définies sur

$$[\Omega, p] \text{ tel que : } p(V_1, V_2, \ldots, V_N) = \prod_{i=1\ldots N} p(vi \mid c(vi)) \qquad (2.1)$$

Avec **C(Vi)** l'ensemble des parents de Vi dans le graphe [BOU 05].

2.4. Pourquoi utiliser des réseaux bayésiens ?

Selon le type d'application, l'utilisation pratique d'un réseau bayésien peut être envisagée au même titre que celle d'autres modèles : réseau de neurones, système expert, arbre de décision, modèle d'analyse de données (régression linéaire), arbre de défaillances, modèle logique. Naturellement, le choix de la méthode fait intervenir différents critères, comme la facilité, le coût et le délai de mise en œuvre d'une solution. En dehors de toute considération théorique, les aspects suivants des réseaux bayésiens les rendent dans de nombreux cas, préférables à d'autres modèles :

- **Acquisition des connaissances**.

  La possibilité de rassembler et de fusionner des connaissances de diverses natures dans un même modèle : retour d'expérience (données historiques ou empiriques), expertise (exprimée sous forme de règles logiques, d'équations, de statistiques ou de probabilités subjectives), observations. Dans le monde industriel, par exemple, chacune de ces sources d'information, quoique présente, est souvent insuffisante individuellement pour fournir une représentation précise et réaliste du système analysé.

- **Représentation des connaissances**.

La représentation graphique d'un réseau bayésien est explicite, intuitive et compréhensible par un non-spécialiste, ce qui facilite à la fois la validation du modèle, ses évolutions éventuelles et surtout son utilisation. Typiquement, un décideur est beaucoup plus enclin à s'appuyer sur un modèle dont il comprend le fonctionnement qu'à faire confiance à une boîte noire.

- **Utilisation des connaissances**.

Un réseau bayésien est polyvalent : on peut se servir du même modèle pour évaluer, prévoir, diagnostiquer, ou optimiser des décisions, ce qui contribue à rentabiliser l'effort de construction du réseau bayésien.

- **Qualité de l'offre en matière de logiciels**.

Il existe aujourd'hui de nombreux logiciels pour saisir et traiter des réseaux bayésiens. Ces outils présentent des fonctionnalités plus ou moins évoluées : apprentissage des probabil

ités, apprentissage de la structure du réseau bayésien, possibilité d'intégrer des variables continues, des variables d'utilité et de décision, etc. Nous allons à présent étudier plus en détail ces différents aspects de l'utilisation de réseaux bayésien [NAï07].

2.5. Représentation graphique de la causalité

Un réseau bayésien (réseau probabiliste ou Bayesian Network) est un modèle représentant des connaissances incertaines sur un phénomène complexe, et permettant, à partir des données, un véritable raisonnement. Un réseau bayésien a pour objectif d'acquérir, de représenter et d'utiliser la connaissance. Il est constitué de deux composantes [BRO 05] :

- Un graphe causal, orienté, acycliques, dont les nœuds sont des variables
- D'intérêt du domaine, les arcs des relations de dépendance entre ces variables. L'ensemble des nœuds et des arcs forme ce que l'on appelle la structure du réseau bayésien. C'est la représentation qualitative de la connaissance.
- Un ensemble de distributions locales de probabilités qui sont les paramètres du réseau. Pour chaque nœud on dispose d'une table de probabilités P(variable/parents(variable)) qui représente la distribution locale de probabilité. Il faut remarquer que l'état de chaque nœud ne dépend que de

l'état de ses parents. Il s'agit de la représentation quantitative de la connaissance.

On peut décrire un réseau bayésien comme un système expert probabiliste. Dans un réseau bayésien, un arc de A vers B peut être interprété par ' A cause B', les cycles ne sont pas autorisés, et le graphe est un graphe acyclique orienté. De plus un nœud est conditionnellement indépendant de ses non-descendants sachant ses parents.

Exemple

Dans les réseaux bayésiens, deux nœuds qui représentent deux faits différents peuvent être en relation causale sans que l'un implique l'autre. Graphiquement, chaque parent a un effet sur ses fils. La notion de causalité joue un rôle très important pour construire les réseaux bayésiens [COR 02].

Exemple de réseau bayésien modélisant l'appel potentiel de voisins, suite au déclenchement d'une alarme possiblement causé par un cambriolage ou un tremblement de terre. Notation abrégée : T = Tremblement De Terre, C = Cambriolage, A = Alarme,
J = JohnAppelle, M = MarieAppelle. Exemple tiré du livre "Artificial Intelligence : A Modern Approach" de Rusell et Norvig [FAT 07].

La représentation graphique du modèle causal utilisé est dans la figure (Fig.2.1). Cette figure représente un réseau bayésien simple contenant cinq variables binaires, Etant donné un réseau bayésien, la distribution conjointe peut être simplifiée comme suit :

$$P(X_1, X2, \ldots, X_n) = P(X_1 | X2, \ldots, Xn)\, P(X_2, \ldots, Xn) = P(X_1 | Parents(X_1))\, P(X_2, \ldots, Xn)$$

Règle du produit indépendance

Ou le terme conjoint signifie qu'on exprime la probabilité d'observer plusieurs variables simultanément.

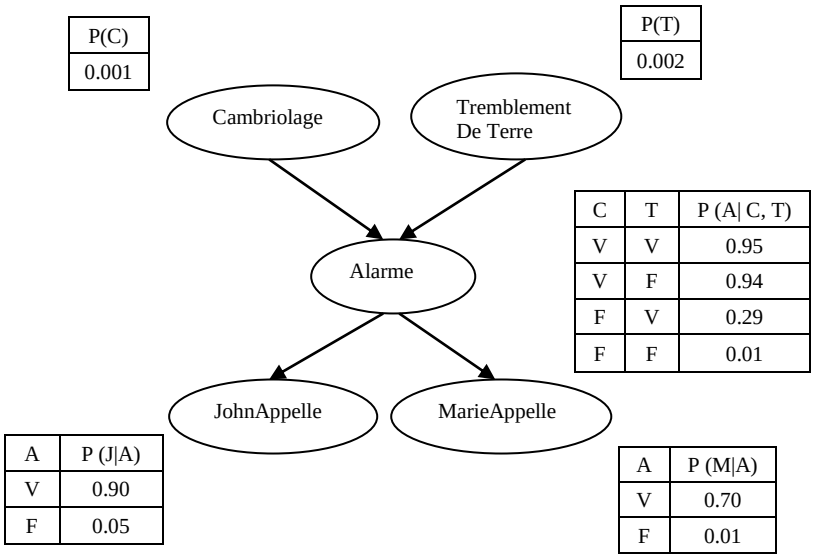


Fig.2. 1 Graphe de causalité

Suivant l'idée présentée plus haut, la distribution conjointe de ce modèle présenté par la (Fig.2.1) peut s'exprimer comme :

$$P(M, J, A, C, T) = P(M|A)\, P(J|A)\, P(A|C, T)\, P(C)\, P(T) \qquad (2.2)$$

2.5.1. Circulation de l'information dans un graphe causal

Nous allons à présent étudier de plus près comment l'information circule au sein d'un *graphe causal*.

À l'aide d'un exemple extrêmement classique dans la littérature sur les réseaux Bayésiens extrait de [PEA 86], et repris dans [JEN 96], nous allons présenter la circulation de l'information dans un graphe causal :

" *Ce matin-là, alors que le temps est clair et sec, M. Holmes sort de sa maison, et s'aperçoit que la pelouse de son jardin est humide. Il se demande alors s'il a plu pendant la nuit, ou s'il a simplement oublié de débrancher son arroseur automatique. Il jette alors un coup d'œil à la pelouse de son voisin, M. Watson, et s'aperçoit qu'elle est également humide. Il en déduit alors qu'il a probablement plu, et décide de partir au travail sans vérifier son arroseur automatique* ".

Soient A, P, J et V des faits tels que :

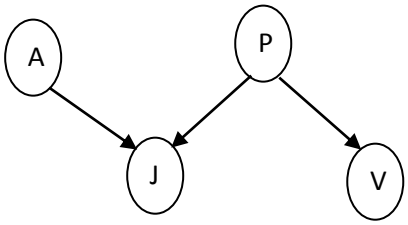
A : j'ai oublié de débrancher mon arroseur automatique.

P : il a plu pendant cette nuit.

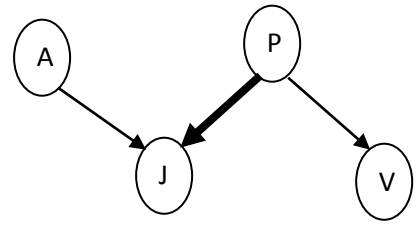
J : l'herbe de mon jardin est humide.

V : l'herbe du jardin de mon voisin est humide.

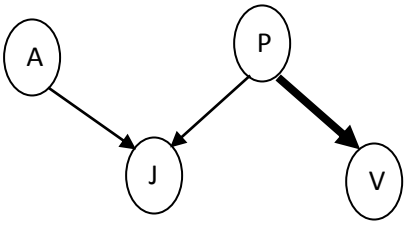
La représentation graphique du modèle causal utilisé par M. Holmes est la suivante :



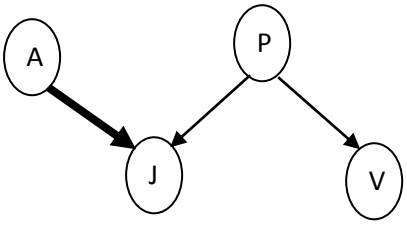
Le raisonnement de M. Holmes se résume dans l'explication suivante des causalités :



S'il a plu pendant la nuit, l'herbe de mon jardin est humide.



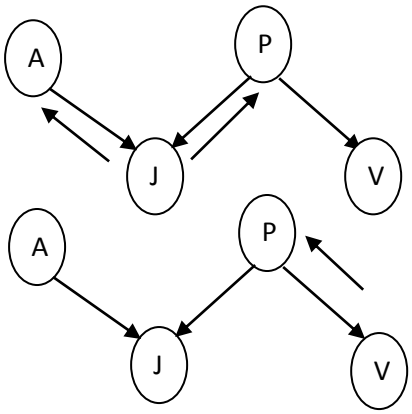
S'il a plu pendant la nuit, l'herbe du jardin de mon voisin est également humide.



Si j'ai oublié de débrancher mon arroseur automatique, l'herbe de mon jardin est humide.

Étant donné que l'information J est vraie, le modèle nous indique que J, a dû être causé soit par A, soit par P. Les deux causes sont a priori plausibles (probables) par manque d'information complémentaire.

La véracité de V renforce la croyance en P, quoique le voisin pourrait lui aussi avoir oublié de débrancher son arroseur automatique. Dans cette première analyse, seulement le sens de circulation Effet → Cause a été considéré.



La connaissance de J renforce la croyance dans l'une des deux causes A ou P

La connaissance de V augmente la croyance dans la cause P. La cause A devient moins plausible

M. Holmes a donc déduit que son arroseur automatique était à l'arrêt à partir du fait que la pelouse de son voisin était humide.

En général, pour expliquer la circulation de l'information dans un graphe causal, Patrick Naïm dans [Naï07] a considéré les 3 cas suivants :

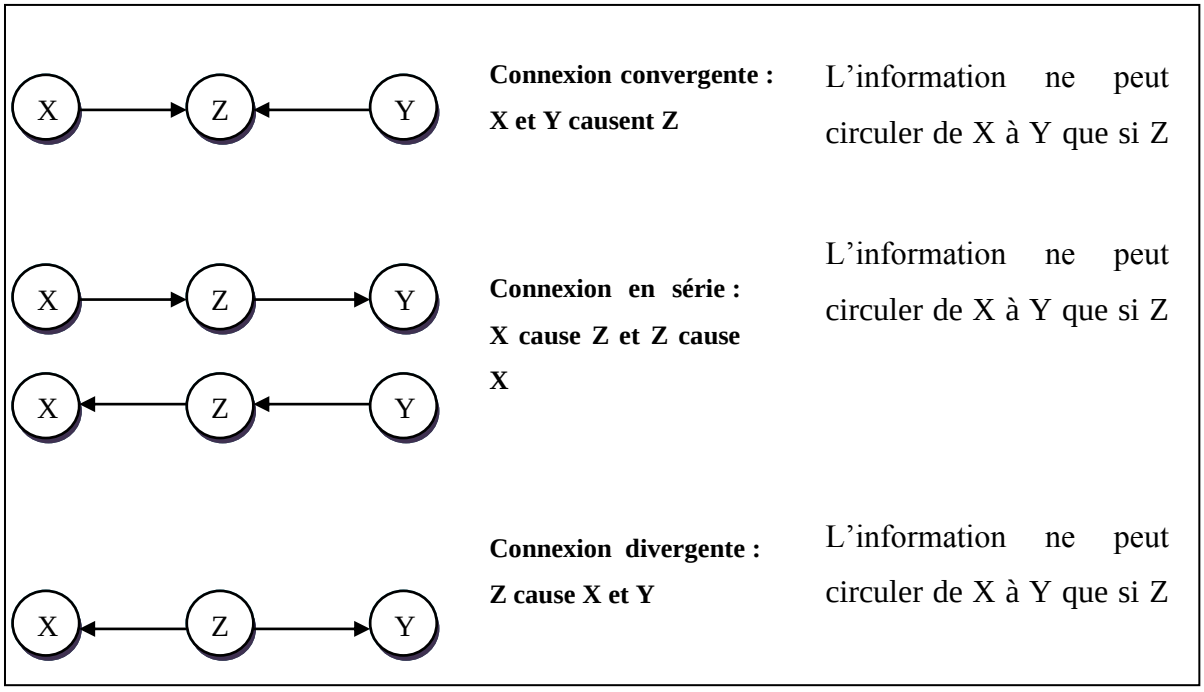


Fig.2. 2 Mode de connexion et Circulation de l'information

2.5.2. Indépendance conditionnelle dans un réseau bayésien : d-séparation

Considérons trois ensembles disjoints de variables *X*, *Y* et *Z* représentés par trois ensembles de nœuds dans un graphe acyclique dirigé *G*. Pour savoir si *X* est indépendant de *Y* sachant *Z* dans toute distribution compatible avec *G*, nous avons besoin de tester si des nœuds

correspondants aux variables de *Z* **bloquent** tous les chemins allant des nœuds de *X* aux nœuds de *Y*.

Un chemin est une séquence consécutive d'arcs (non-dirigés) dans le graphe.

Un blocage peut être vu comme un arrêt du flux d'informations entre les variables qui sont ainsi connectées.

Le flux d'information est dirigé par le sens des arcs et représente le flux des causalités dans le graphe, ou l'ordre dans lequel les influences vont se propager dans le graphe. Cette propagation des influences peut alors être vue comme un envoi d'information d'une variable à ses variables filles [GEI 90].

2.5.2.1. Définition d-Séparation

C'est un critère général pour décider si un nœud X est indépendant d'un nœud Y, étant donnés d'autres nœuds Z= $\{Z_{y,\ldots,}$ Zm$\}$

Soit un réseau bayésien de structure D = (V, E) et X, Y, Z trois sous-ensembles disjoints de variables (et donc de nœuds) de D.

- On dit que X et Y sont ***d- séparé*** par Z si toute chaine reliant X et Y et ***bloquées*** par Z
- On dit que X est ***indépendant*** de Y sachant Z si les chemins non-dirigés (c à d tous les nœuds entre X et Y qui ne respectent pas nécessairement le sens de flèche) entre X et Y sont ***bloqués*** par Z.
- Si les ensembles X et Y sont ***d-séparés*** par Z dans un GAD G, alors X est indépendant de Y conditionnellement à Z. Réciproquement, si X et Y ***ne sont pas d-séparés*** par Z dans un GAD G, alors X et Y sont dépendants conditionnellement à Z.
- ***Un chemin est bloqué*** s'il contient au moins un nœud *N* qui satisfait une ou l'autre des conditions suivantes :
    1. Il inclue un nœud → (N) → (connexion série) ou ← (N) → (connexion divergente), ou Z $\in\{Z_{y\ldots,}$ Zm$\}$ c à d si *N* est observé.
    2. Il inclue un nœud → (N) ← (connexion convergente), $N \notin \{Z_{y\ldots,}$ Zm$\}$ (*N* n'appartiens pas aux variables observées ni aucun de ces descendants)

Présentons 3 aspects des règles de d-séparation et l'indépendance conditionnelle sur le graphe dans les figures ((fig.2.3), (fig.2.4), (fig.2.5)) suivantes :

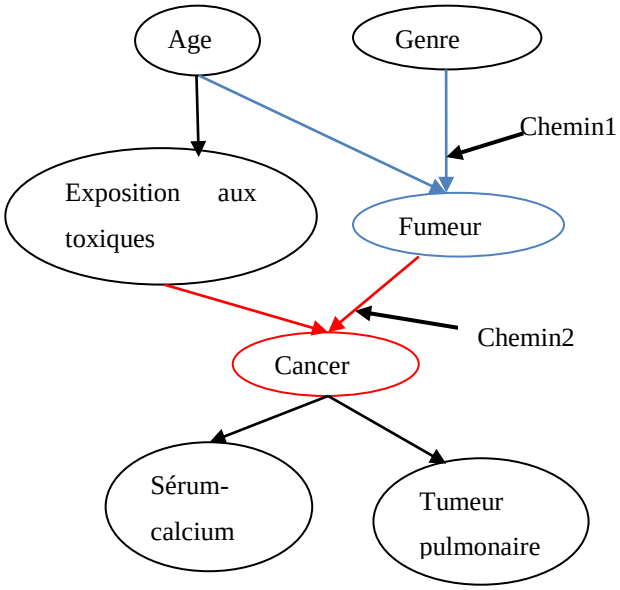
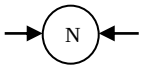
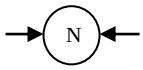


Fig.2. 3 Exemple1 de d-séparation

- Est-ce que Age et Gender sont indépendants ?
  - Le chemin 1 est bloqué au niveau de Fumeur car on une connexion convergente  .

-En plus Fumeur est ces descendants : cancer, sérum-calcium et Tumeur pulmonaire *ne sont pas observés* Z= {}

  - Le chemin 2 est aussi bloqué au niveau de cancer pour les mêmes raisons 

Donc Age et Genre ***sont indépendants***.

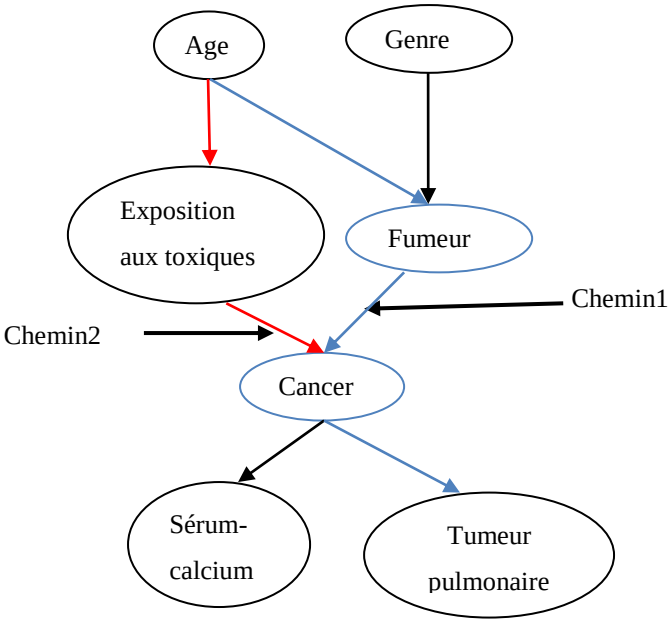
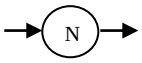
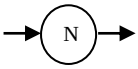
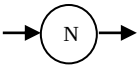


Fig.2. 4 Exemple2 de d-séparation

- Est-ce que Age et Tumeur pulmonaire sont indépendants sachant Fumeur ?
  - ➢ Le chemin 1 est bloqué au niveau de Fumeur 

    Mais *Fumeur est observé*
  - ➢ Le chemin 2 n'est pas bloqué.
    - Exposition aux toxiques 

      n'est pas observée.
    - Cancer 

      n'est pas observé.
- Donc Age et Tumeur pulmonaire ***ne sont indépendants pas.***

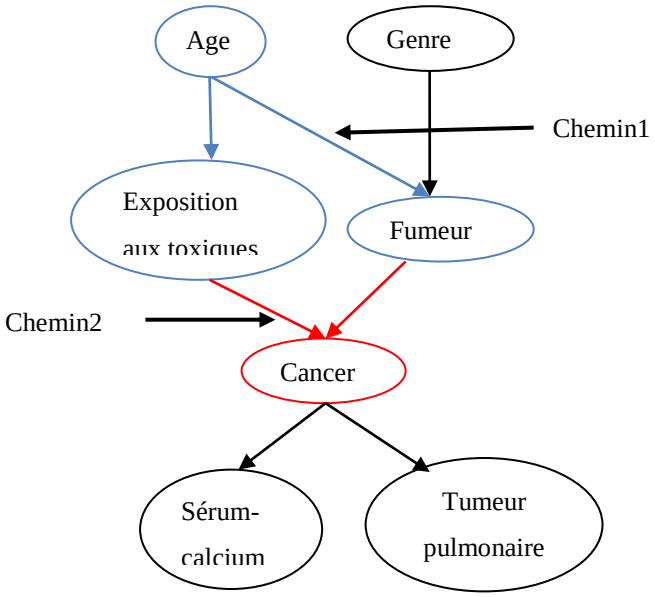
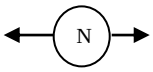
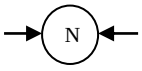


Fig.2. 5 exemple3 de d-séparation

Est-ce qu'Exposition aux toxiques et Fumeur sont indépendants sachant Age et Tumeur pulmonaire ?

- ➢ Le chemin 1 est bloqué au niveau de l'Age 

  Mais Age *est observé*
- ➢ Le chemin 2 n'est pas bloqué.
  -Cancer 

  ne bloque pas le chemin puisque Tumeur pulmonaire, un de ses descendants, est observé.
- Donc Exposition aux toxiques et Fumeur ***ne sont indépendants pas.***

2.6. Une représentation probabiliste associée

Avec la représentation graphique de la causalité on peut connaître la direction de circulation de connaissances dans le graphe mais on ne peut pas connaître la quantité de

circulation de connaissances. Alors, il faut une représentation probabiliste associé avec le graphe. Avec une relation causale : A=>B on peut représenter la quantité de cette relation par la probabilité conditionnelle : p(B|A).

2.6.1. L'indépendance conditionnelle

- Définition :

A et B sont indépendantes conditionnellement à C si seulement si :

- Lorsque l'état de C est connu toute connaissance sur B n'altère pas A.
- P (A|B, C) =P(A|C).

-Les réseaux bayésiens permettent de représenter graphiquement les indépendances conditionnelles

- 3 types de relations sont possibles entre A, B et C

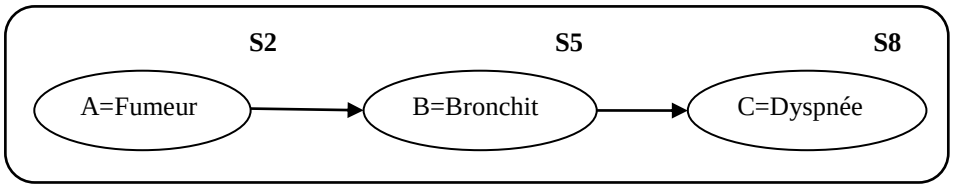


Fig.2. 6 Exemple de connexion en série

- Connexion en série décrit dans la (Fig.2.6)
- A et B sont dépendants.
- A et B sont indépendantes conditionnellement à C si seulement si :
  - P (C) est connue {P(A) n'intervient pas le calcule de P(B)}.
  - P (S8|S5, S2) =P (S8|S4) =P (S8|Parents (S8)).

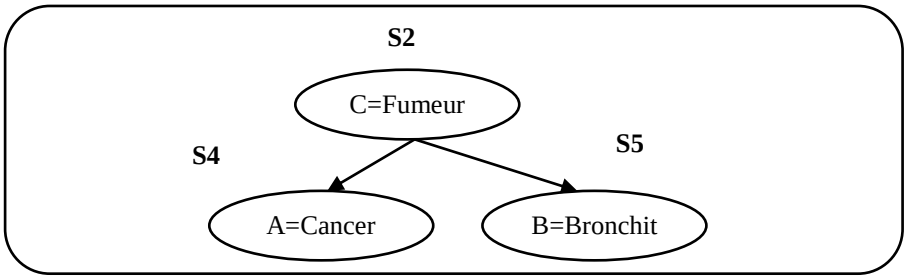


Fig.2. 7 Exemple de connexion divergente

- Connexion divergente décrit dans la (Fig.2.7)
- A et B sont dépendants.

- A et B sont indépendantes conditionnellement à C si seulement si :
  - ❖ P (C) est connue {P(A) n'intervient pas le calcule de P(B)}.
  - ❖ P (S5|S2, S4) =P (S5|S2) =P (S4|Parents (S4)).

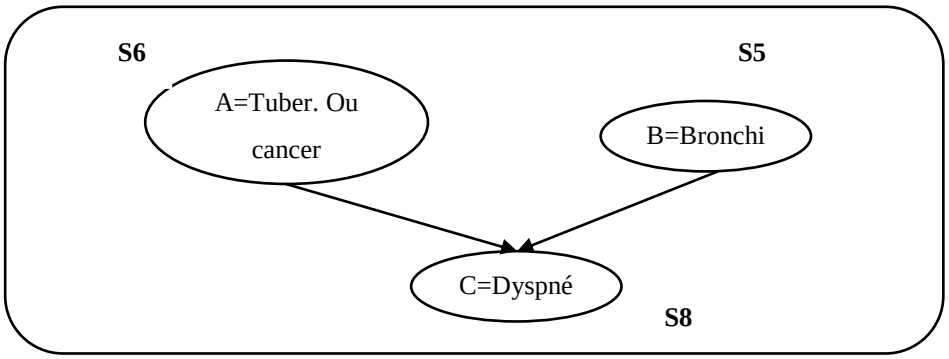


Fig.2. 8 Exemple de connexion convergente

- Connexion convergente décrit dans la (Fig.2.8)
- A et B sont indépendant s.
- A et B sont indépendantes conditionnellement à C si seulement si :
  - ❖ P (C) est connue {P(A) n'intervient pas le calcule de P(B)}
  - ❖ P (S8|S6, S5) =P (S8|Parents (S8)) [LER 97].

2.7. Formule de Bayes

Le théorème de Bayes est utilisé dans l'inférence statistique pour mettre à jour ou actualiser les estimations d'une probabilité ou d'un paramètre quelconque, à partir des observations et des lois de probabilité de ces observations. Il existe une version *discrète* et une version continue du théorème [FIN 96].

- L'école *bayésienne* utilise les probabilités comme moyen de traduire numériquement un degré de connaissance (la théorie mathématique des probabilités n'oblige en effet nullement à associer celles-ci à des fréquences, qui n'en représentent qu'une application particulière résultant de la loi des grands nombres). Dans cette optique, le théorème de Bayes peut s'appliquer à toute proposition, quelle que soit la nature des variables et indépendamment de toute considération ontologique.
- L'école *fréquentiste* utilise les propriétés de long terme de la loi des observations et ne considère pas de loi sur les paramètres, inconnus mais fixés.

En théorie des probabilités, le théorème de Bayes énonce des probabilités conditionnelles : Etant donné deux évènements *A* et *B*, le théorème de Bayes permet de déterminer la probabilité de *A sachant B*, si l'on connaît les probabilités :

- De *A* ;
- De *B* ;
- De *B sachant A*.

Ce théorème élémentaire originellement nommé « de probabilité des causes » a des applications considérables.

Pour aboutir au théorème de Bayes, on part d'une des définitions de la probabilité conditionnelle [FIN 96] :

$$\mathcal{P}(A|B)\ \mathcal{P}(B) = \mathcal{P}(A \cap B) = \mathcal{P}(B|A)\ \mathcal{P}(A) \qquad (2.3)$$

En notant $\mathcal{P}(A|B)$ la probabilité que *A* et B aient tous les deux lieux. En divisant de part et d'autre par $\mathcal{P}(B)$, on obtient :

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)} \qquad (2.4)$$

Soit le théorème de Bayes. Chaque terme du théorème de Bayes a une dénomination usuelle. Le terme *P*(*A*) est la probabilité a priori de *A*. Elle est « antérieure » au sens qu'elle précède toute information sur *B*. *P*(*A*) est aussi appelée la probabilité marginal*e* de *A*. Le terme *P*(*A*|*B*) est appelé la probabilité a posteriori de *A* sachant *B* (ou encore de *A* sous condition *B*). Elle est « postérieure », au sens qu'elle dépend directement de *B*. Le terme *P*(*B*|*A*), pour un *B* connu, est appelée la fonction de vraisemblance de *A*. De même, le terme *P*(*B*) est appelé la probabilité marginale ou *a* priori de *B*.

2.7.1. Autres écritures du théorème de Bayes

On améliore parfois le théorème de Bayes en remarquant que [FIN 96] :

$$P(B) = P(A \cap B) + P(A^C \cap B) = P(B|A)\ P(A) + P(B|A^C)\ P(A^C) \qquad (2.5)$$

Afin de réécrire le théorème ainsi :

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B|A)P(A) + P(B|A)P(B|A^C)\ P(A^C)} \qquad (2.6)$$

Où A^C est le complémentaire de *A*. Plus généralement, si $\{A_i\}$ est une partition de l'ensemble des possibles,

$$P(A_i|B) = \frac{P(B|A_i)\,P(A_i)}{\sum_j P(B|A_j)\,P(A_j)} \qquad (2.7)$$

2.7.2. Exemple d'application de la formule de Bayes

Dans une population pour laquelle 1 habitant sur 100 est atteint d'une maladie génétique A, on a mis au point un test de dépistage. Le résultat du test est soit positif (T) soit négatif ($\overline{T}$)

On sait que :

P(T |A)=0,8 et P ($\overline{T}|\overline{A}$)=0,9

On soumet un patient au test. Celui-ci est positif. Quelle est la probabilité que ce patient soit atteint de la maladie A soit P(A) ou P (A|T) ?

D'après la formule de Bayes :

$$P(A|T) = \frac{P(A \cap T)}{P(T)} = \frac{P(T|A)\,P(A)}{P(T|A)P(A) + P(T|\overline{A})P(\overline{A})}$$

D'où

$$P(A|T) = \frac{0{,}01 * 0{,}8}{0{,}8 * 0{,}01 + 0{,}1 * 0{,}99} = 0{,}075$$

Ainsi *avant le test*, la probabilité d'être malade était de P(A) = 0 ,01 (probabilité a priori) et *après le test* la probabilité d'être malade est de P (A/T) = 0,075 (probabilité a posteriori). Ainsi le test apporte un supplément d'information.

La formule de Bayes est utilisée de façon classique pour calculer des probabilités de causes dans des diagnostics (maladies, pannes, etc.).

L'application du théorème de Bayes est à la base de toute une branche de la statistique appelée *statistique bayésienne.*

Le nom du réseau bayésien donc provient de la formule d'inversion de bayes :

$$P(A|B) = \frac{P\,(B|A)\,P\,(A)}{P\,(B)}$$

Avec :

(A) : l'hypothèse.

(B) : l'observation.

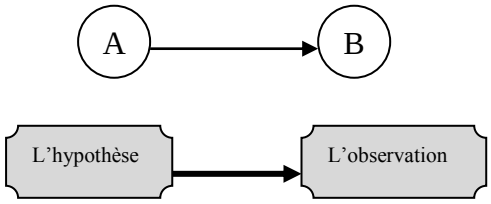


Fig.2. 9 règle de la vraisemblance

On définit alors :

P(A|B)	La probabilité à posteriori
P(B)	Probabilité d'observer les données B indépendamment de l'hypothèse (évidence)
P(A)	Probabilité que l'hypothèse A soit vérifiée indépendamment des données de B (probabilité a priori)
P(B|A)	Probabilité d'observer les données B sachant que l'hypothèse A est vérifiée (vraisemblance)

Le théorème de bayes exprime le fait que si l'évènement B est observé, notre hypothèse sur l'évènement A doit être par la loi à posteriori p(A|B) qui est obtenue en multipliant la probabilité à priori P(A) par la probabilité p(B|A) qu'on appelle vraisemblance.

probabilité à posteriori= vraisemblance * probabilité à priori

2.8. Construction des réseaux bayésiens :

Le réseau bayésien peut être construit :

- Par expertise,
- Par analyse fonctionnelle,
- Par apprentissage, en exploitant une base de données (ce procédé est encore au stade de recherche et développement).

La construction d'un réseau bayésien s'effectue en trois étapes essentielles, qui sont présentées sur la (Fig.2.9) ci-après. Chacune des trois étapes peut impliquer un recueil d'expertise, au moyen de questionnaires écrits, d'entretiens individuels ou encore de séances de *brainstorming* [NAÏ 07].

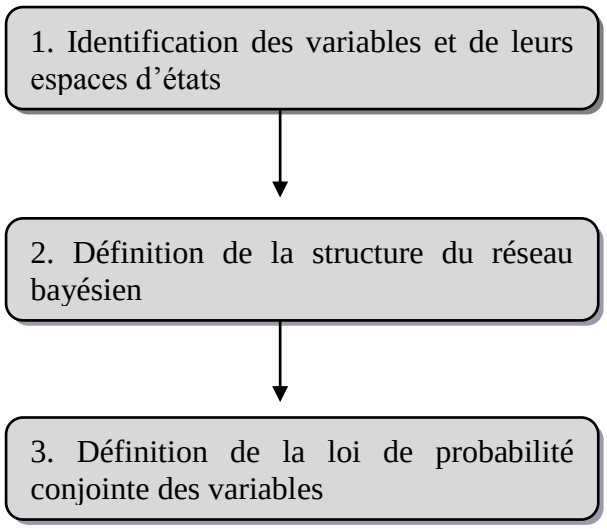


Fig.2. 10 Étapes de construction d'un réseau bayésien

2.8.1. Identification des variables et de leurs espaces d'états

La première étape de construction du réseau bayésien est la seule pour laquelle l'intervention humaine est absolument indispensable. Il s'agit de déterminer l'ensemble des variables Xi, catégorielles ou numériques, qui caractérisent le système. Comme dans tout travail de modélisation, un compromis entre la précision de la représentation et la maniabilité du modèle doit être trouvé, au moyen d'une discussion entre les experts et le modélisateur.
Lorsque les variables sont identifiées, il est ensuite nécessaire de préciser l'*espace d'états* de chaque variable Xi, c'est-à-dire l'ensemble de ses valeurs possibles.

La majorité des logiciels de réseaux bayésiens ne traite que des modèles à variables discrètes, ayant un nombre fini de valeurs possibles. Si tel est le cas, il est impératif de discrétiser les plages de variation des variables continues. Cette limitation est parfois gênante en pratique, car des discrétisations trop fines peuvent conduire à des tables de probabilités de grande taille, de nature à saturer la mémoire de l'ordinateur [NAÏ 07].

2.8.2. Définition de la structure du réseau bayésien

Une fois les variables aléatoires identifiées, il faut définir les dépendances (les influences) qui les relient.

Chaque variable :

- Influence d'autres variables,
- Est influencée par d'autres variables.

En d'autre terme la deuxième étape consiste à identifier les liens entre variables, c'est-à-dire à répondre à la question : pour quels couples (i, j) la variable Xi influence-t-elle la variable Xj?

Dans la plupart des applications, cette étape s'effectue par l'interrogation d'experts. Dans ce cas, des itérations sont souvent nécessaires pour aboutir à une description consensuelle des interactions entre les variables Xi. L'expérience montre cependant que la représentation graphique du réseau bayésien est dans cette étape un support de dialogue extrêmement précieux [NAÏ 07].

2.8.3. Loi de probabilité conjointe des variables

La dernière étape de construction du réseau bayésien consiste à renseigner les tables de probabilités associées aux différentes variables. Dans un premier temps, la connaissance des experts concernant les lois de probabilité des variables est intégrée au modèle. Concrètement, deux cas se présentent selon la position d'une variable Xi dans le réseau bayésien :

- La variable Xi n'a pas de variable parente : les experts doivent préciser la loi de probabilité marginale de Xi.
- La variable Xi possède des variables parentes : les experts doivent exprimer la dépendance de Xi en fonction des variables parentes, soit au moyen de probabilités conditionnelles, soit par une équation déterministe (que le logiciel convertira ensuite en probabilités).

Le recueil de lois de probabilités auprès d'experts est une étape délicate du processus de construction du réseau bayésien. Typiquement, les experts se montrent réticents à chiffrer la plausibilité d'un événement qu'ils n'ont jamais observé.

Cependant, une discussion approfondie avec les experts, aboutissant parfois à une reformulation plus précise des variables, permet dans de nombreux cas l'obtention d'appréciations qualitatives. Ainsi, lorsqu'un événement est clairement défini, les experts sont généralement mieux à même d'exprimer si celui-ci est probable, peu probable, hautement improbable, etc.

Le cas d'absence totale d'information concernant la loi de probabilité d'une variable Xi peut être rencontré. La solution pragmatique consiste alors à affecter à Xi une loi de probabilité arbitraire, par exemple une loi uniforme. Lorsque la construction du réseau bayésien est achevée, l'étude de la sensibilité du modèle à cette loi permet de décider ou non de consacrer davantage de moyens à l'étude de la variable Xi.

La quasi-totalité des logiciels commerciaux de réseaux bayésiens permet l'apprentissage automatique des tables de probabilités à partir de données. Par conséquent, dans un second temps, les éventuelles observations des Xi peuvent être incorporées au modèle, afin d'affiner les probabilités introduites par les experts [NAÏ 07].

2.9. Inférence bayésienne

Définition 1

L'inférence dans un réseau bayésien concerne le calcul de la probabilité de n'importe quelle variable ou sous ensemble de variables à partir des autres variables observées. Il s'agit donc de déterminer les probabilités conditionnelles d'évènements reliés par des relations d'influence [PEA88].

Définition 2

L'inférence dans un réseau bayésien se résume à un calcul de probabilités a posteriori. Connaissant les états de certaines variables (appelées variables d'observation), on détermine les probabilités des états de certaines autres variables (appelées variables cibles) conditionnellement aux observations [LAU 88].

Définition 3

L'inférence est le fait de propager les informations des nouvelles évidences du réseau pour pouvoir calculer de quelle manière elles influent sur les autres variables du système et ainsi nous permettre de connaître avec moins d'à priori l'état du système observé [CAN 04]. L'inférence bayésienne est définie par [NAÏ 07] comme le processus de propager une ou plusieurs informations certaines au sein d'un réseau pour en déduire comment sont modifiées les croyances concernant les autres nœuds. En d'autres termes, l'inférence sert calculer la probabilité d'une hypothèse suite l'observation des évidences. Les évidences correspondent aux nœuds d'entrée et les hypothèses sont les différents états des nœuds de sortie du réseau.

L'injection des probabilités des nœuds d'entré va modifier récursivement les probabilités des nœuds enfants jusqu'aux nœuds de sortie. Le calcul des probabilités utilise la fois les tables de probabilités et le théorème de Bayes [ZOG 11].

2.9.1. Les algorithmes d'inférence

Il existe plusieurs algorithmes d'inférence dans les réseaux bayésiens class s en deux groupes [BOU 12]. D'un côté nous avons les méthodes l'inférence exactes qui exploitent les indépendances conditionnelles contenues dans les réseaux et donnent à chaque inférence les probabilités à posteriori exactes. Par exemple l'algorithme Clustering [GUO 04] effectue l'inférence en transformant le réseau en un arbre pour lequel chaque nœud regroupe plusieurs nœuds du réseau initial. D'autre côté nous avons les méthodes approchées qui estiment les probabilités a posteriori. Pour ces méthodes, deux exécutions d'une inférence peuvent donner des probabilités à posteriori différentes [FUN 13a]. Likelihood weighting [FUN 13], Backwardsampling[FUN 13a], Self importance et Heuristic importance [SHA 13]) qui estiment les probabilités en effectuant plusieurs tirages dans l'ensemble des combinaisons possibles des états des variables du réseau.

2.9.1.1. Méthodes d'inférence exactes

1) Messages locaux

La première méthode d'inférence, est celle des messages locaux, plus connue sous le nom « polytree algorithm ». Elle consiste en une actualisation, à tout moment, des probabilités marginales, par transmission de messages entre variables voisines dans le graphe d'indépendance. Cette méthode ne fonctionne de manière exacte que lorsque le réseau bayésien possède une forme d'arbre (ou polytree en anglais) [KIM 83].

2) Ensemble de coupe

L'algorithme Loop Cutset Conditioning a été introduit très tôt par Pearl dans cette méthode, la connectivité du réseau est changée en instanciant un certain sous ensemble de variables appelé l'ensemble de coupe (loop cutset). Dans le réseau résultant, l'inférence est effectuée en utilisant l'algorithme des messages locaux. Puis les résultats de toutes les instanciations sont combinés par leurs probabilités a priori. La complexité de cet algorithme augmente donc exponentiellement en fonction de la taille de l'ensemble de coupe [PEA 86].

3) Arbre de jonction

La méthode de l'arbre de jonction (aussi appelée clustering ou clique-tree propagation algorithm) a été introduite par Lauritzen & Spiegelhalter [LAU 88] et Jensen, Lauritzen& Olesen [JEN 90]. Elle est aussi appelée méthode JLO (pour Jensen, Lauritzen, Olessen). Elle est applicable pour toute structure de GOA contrairement à la méthode des messages locaux. Néanmoins, s'il y a peu de circuits dans le graphe, il peut très préférable d'utiliser une méthode basée sur un ensemble de coupe. Cette méthode est divisée en cinq étapes qui sont :

- Moralisation du graphe,
- Triangulation du graphe moral,
- Construction de l'arbre de jonction,
- Inférence dans l'arbre de jonction en utilisant l'algorithme des messages locaux,
- Transformation des potentiels de clique en lois conditionnelles mises à jour.

4) Elimination de variables

L'élimination de variables est décrite dans Zhang & Poole [ZHA 94]. Cet algorithme supprime les variables une par une après avoir sommé sur celles-ci. Cette méthode a été généralisée dans Dechter [DEC 90] par l'algorithme Bucket Elimination. Un ordre des variables doit être donné en entrée et sera alors l'ordre d'élimination des variables. Le nombre de calculs dépend alors de cet ordre puisqu'il influe sur la taille des facteurs futurs. Trouver le meilleur ordre qui vaut au problème de trouver l'arbre de plus petite largeur dans le réseau ce qui est un problème NP-dur. Cette méthode est avantageuse lorsqu'un ordre d'élimination des variables est déjà connu ou si le réseau est peu dense mais avec de nombreux circuits.

5) Explication la plus probable

La méthode de l'explication la plus probable (MPE pour Most probable explanation) n'est pas réellement une technique d'inférence mais plutôt un problème d'inférence. Ce problème consiste en l'identification de l'état le plus probable. Il est possible d'adapter différentes méthodes d'apprentissage pour répondre à cette question. La technique la plus commune (et exacte, Lauritzen & Spiegelhalter [LAU 88] pour effectuer cette inférence consiste en le remplacement des signes sommes par des max et les signes produits par des min dans les formules de l'inférence classique. Il est possible d'adapter cette méthode pour trouver le deuxième cas le plus probable ou, plus généralement, le n-ième cas le plus probable. Comme pour les autres problèmes d'inférence, il existe également des algorithmes approchés

pour résoudre ce problème : par exemple, Guo, Boddhireddy& Hsu (2004) [GUO 04] propose une méthode à base de colonies de fourmis.

6) Méthodes symboliques

L'inférence probabiliste symbolique (SPI pour Symbolic Probabilistic Inference) a été introduite dans Shachter, D'Ambrosio, & Del Fabero [SHA 90] et Li & D'Ambrosio [ZHA 94]. Cette méthode est orientée par un but : n'effectuer que les calculs nécessaires pour répondre à la requête. Des expressions symboliques peuvent être obtenues en remettant à plus tard l'évaluation des expressions, et en les gardant sous forme symbolique.

Par ailleurs, Castilo, Guti rez, & Hadi [CAS 96] ont proposés une autre technique d'inférence symbolique en modifiant les méthodes existantes d'inférences numériques et en remplaçant les paramètres initiaux par des paramètres symboliques.

Ces méthodes ont le désavantage qu'il est difficile de calculer et de simplifier automatiquement des expressions symboliques mais, l'avantage qu'elles orientent les calculs.

7) Méthodes différentielles

Les méthodes différentielles transforment un réseau bayésien en polynôme multi varié (Darwiche) [DAR 00]. Elles calculent ensuite les dérivées partielles de ce polynôme. Il est alors possible d'utiliser ces dérivées pour calculer les réponses à de nombreuses requêtes, et cela en temps constant. Cette méthode est très utile lorsque nous effectuons régulièrement les mêmes requêtes car maintenant nous pourrons y répondre en temps constant.

2.9.1.2. L'inférence approximative

Lorsqu'il n'est pas possible de faire une inférence exacte, ce qui est le cas des réseaux bayésiens où il y a beaucoup de cycle et/ou de parents par nœud. Trois méthodes différentes ont été proposées. Il s'agit des méthodes :

- *Variationnelles* : où on introduit des paramètres dits variationnelles qui permettent de briser les cycles du réseau. Pour faire de l'inférence, il suffit de résoudre un problème d'optimisation, minimiser la distance de Kullback-Leibler. D'une manière intuitive, cela revient à trouver la meilleure approximation de la distribution P(E), trouver la meilleure distribution sur les variables cachées [JEN 96].

- *De Monte-Carlo* : où on génère une série d'échantillon qui vont permettre d'estimer les distributions de probabilités. Les méthodes de Monte Carlo ont pour but e résoudre un des deux problèmes suivants [MAC 96].

-Générer des échantillons qui suivent une loi de probabilité P(X = x) donnée.

-Estimer des espérances de fonctions suivant cette distribution.

- *De propagation cyclique* : où on utilise des techniques utilisées pour les réseaux sans cycles (Propagation cyclique) Le passage de message (propagation) est effectué avec l'algorithme de Pearl.

2.10. Apprentissage

On a déjà vu qu'un réseau bayésien est constitué à la fois d'un graphe et d'un ensemble de probabilités conditionnelles. L'apprentissage d'un réseau bayésien doit donc répondre aux deux questions suivantes :

- Comment estimer les lois de probabilités conditionnelles ?
- Comment trouver la structure du réseau bayésien ?

Le problème de l'apprentissage se divise en deux parties :

- L'apprentissage des paramètres, où il faut estimer les probabilités conditionnelles de chaque nœud du réseau.
- L'apprentissage de la structure, où le but est de trouver le meilleur graphe représentant la tâche à résoudre.

2.10.1. Apprentissage des paramètres

2.10.1.1. À partir de données complètes

On va chercher ici à estimer les distributions de probabilités (ou les paramètres des lois correspondantes) à partir de données disponibles. L'estimation de distributions de probabilités, paramétriques ou non, est un sujet très vaste et complexe. On va décrire ici les méthodes les plus utilisées dans le cadre des réseaux bayésiens, selon que les données sont complètes ou non.

Dans le cas où toutes les variables sont observées, la méthode la plus simple et la plus utilisée est l'estimation statistique qui consiste à 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 :

$$P\ (X_i=x_k|parent(X_i)=c_j)=\theta_{i,j,k}=\frac{N_{i,j,k}}{\sum_k N_{i,j,k}} \qquad (2.8)$$

Dans la formule précédente $N_{i,j,k}$ 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 c_j.

2.10.1.2. À partir de données incomplètes

Dans les applications pratiques, les bases de données 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 EM (Expectation Maximisation) proposé par Dempster [DEM 77]

- $X_v= \{X_V^I\}_{i=1\ldots n}$ 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.

$$N_{i,j,k}^* = E[N_{i,j,k}] = \sum_{i=\&} p(X_i = x_k\ |parent(X_i) = c_j\ , X_v^{(t)}\theta^{(t)} \qquad (2.9)$$

Cette étape revient à réaliser une série d'inférences (exactes ou approchées) en utilisant les paramètres courants du réseau, puis 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 devient possible de calculer de nouveaux paramètres
- $\theta^{(t+1)}$ par maximum de vraisemblance

$$\theta_{i,j,k}^{(t+1)} = \frac{N_{i,j,k}^*}{\sum_K N_{i,j,k}^*} \qquad (2.10)$$

L'algorithme EM s'écrit donc :

Initialiser $\theta^{(0)}$, t=0

- Répéter

t=t+1

Calculer $N^{*}_{i,j,k}$

Calculer $\theta^{(t)}_{i,j,k}$

- Tant que$|\theta^{(t)} - \theta^{(t-1)}| > \varepsilon$
- Fin

L'abréviation l'algorithme EM :

E : Estimer les valeurs manquantes à partir des paramètres actuels $\theta^{(t)}$

- Calculer P (X manquant |X mesurés) dans le RB actuel
- Faire des inférences

M : ré-estimer les paramètres (t+1) à partir des données complétées, en utilisant MV : Max de vraisemblance, MAP : Max A Posteriori, ou EAP : Espérance A Posteriori de i, *j*, *k*.

2.10.2. Apprentissage de la structure

En supposant que la structure de ce réseau est déjà connue. Se pose maintenant le problème de l'apprentissage de cette structure. Comment trouver la structure qui représentera le mieux notre problème. Une première approche consiste à rechercher les différentes relations causales qui existent entre les variables. Les autres approches essaient de quantifier l'adéquation d'un réseau bayésien au problème à résoudre, c'est-à-dire d'associer un score à chaque réseau bayésien. Puis elles recherchent la structure qui donnera le meilleur score dans l'espace des graphes acycliques dirigés. Une approche exhaustive est impossible en pratique en raison de la taille de l'espace de recherche. Le nombre de structures possibles à partir de n nœuds est super exponentiel. Pour résoudre ce problème, ont été proposées un certain nombre d'heuristiques de recherche dans l'espace des graphes, qui restreignent cet espace à l'espace des arbres, ordonnent les nœuds pour limiter la recherche des parents possibles pour chaque variable, ou effectuent une recherche gloutonne dans l'espace.

2.11. Avantages et inconvénients des modèles bayésiens :

Selon [BEN 06] les avantages dans les modèles bayésiens se présentent dans les points suivants :

- Il permet de supporter des données bruitées ;
- Il permet de supporter des données manquantes.

- Il associe des probabilités aux prédictions, ce qui est utile dans les nombreux domaines où les connaissances sont incertaines.
- Il est utilisé par certaines techniques de classification qui sont parmi les plus pratiques ou les plus performantes dans certains domaines.
- Il permet le support de connaissances a priori.
- Il fournit une approche théorique quantitative permettant l'analyse d'autres approches qui ne sont pas nécessairement basées sur des modèles probabilistes.
- Il permet le traitement incrémental des données.

Les réseaux bayésiens ont des avantages supplémentaires qui sont liés à leurs représentations

- Leurs représentations permettent le raisonnement de façon bidirectionnelle (en suivant les relations de dépendances entre variables dans les deux directions).
- Leurs représentations facilitent la compréhensibilité dans un domaine de connaissance (elles représentent directement les connaissances du domaine et non des procédures de raisonnement).
- Leurs représentations modélisent explicitement tous les liens de dépendances entre variables.
- La représentation des connaissances est intuitive et facilement compréhensible.
- Ces modèles s'adaptent aux changements en apprenant au fur et à mesure les nouvelles informations.
- Fondés sur des théories solides (graphes et probabilités), ils représentent au mieux la connaissance disponible à un instant donné.

Inconvénients :

- Le raisonnement (et l'apprentissage) bayésien est associé à certains désavantages
- Son application nécessite des probabilités dont la détermination requière typiquement de grandes quantités de données ou plusieurs connaissances a priori.
- Il nécessite un coût de calcul relativement élevé pour déterminer l'hypothèse optimale dans un cas général.
- Un modèle de probabilité n'est parfois pas un concept intuitif pour un expert du domaine.

- Les réseaux bayésiens ont des désavantages supplémentaires qui sont liés à leurs représentations :
- La compréhensibilité des réseaux peut devenir difficile avec plusieurs variables et/ou plusieurs liens de dépendances.
- Les données continues doivent être discrétisées.

2.12. Conclusion

Les réseaux bayésiens sont une manière naturelle de représenter les dépendances causales. C'est une représentation compacte des distributions jointes. Généralement facile à construire.

Dans ce chapitre nous avons également passé en revue différentes déclinaisons de ces derniers. Ceci permet de nous rendre compte que les réseaux bayésiens sont un formalisme unificateur pour différentes modélisations ayant été développées dans la littérature pour des problématiques aussi vastes que la classification, l'extraction d'information, ou encore simplement pour la modélisation.

Les réseaux bayésiens qui tirent leur profile de la théorie des graphes et des probabilités en particulier par l'intégration de la formule de bayes ont la capacité de faire l'actualisation des données. L'intérêt majeur des réseaux bayésiens est le calcul des probabilités à posteriori des variables, en se basant sur les connaissances des experts et l'observation (évidence). En effet, ceci permet de capitaliser les connaissances et d'enrichir la base de données qui va servir plus tard comme données à priori pour le système.

La plupart des algorithmes développés pour l'inférence et l'apprentissage dans les réseaux bayésiens, aussi bien que les outils disponibles sur le marché pour mettre en œuvre ces algorithmes utilisent des variables discrètes.

Le chapitre suivant nous allons aborder une briefe étude sur les fonctionnalités des outils des réseaux bayésiens, les domaines d'application utilisés ainsi nous parlerons sur le contexte du d'incertitude médicale.

Chapitre 3 Logiciels et domaine d'application

3.1. Introduction

Le développement rapide de la recherche sur le réseau bayésien au cours des 15 dernières années a été accompagné d'une prolifération des logiciels sur les réseaux bayésiens. Ces logiciels ont été construits pour soutenir à la fois ces efforts de recherche et les applications des réseaux bayésien dans une gamme toujours plus large de domaines.

Nous décrivons certains des principaux logiciels, ceux avec le plus de fonctionnalités ou avec une particularité d'intérêt aussi d'après les projets de recherche ou le développement d'applications, nous citons : Hugin, BNJ, BayesiaLab, GeNie. Dans cette analyse des ressources, nous ignorons généralement les aspects de l'interface graphique (Options de menu, icônes abrégées, glisser-déposer, aide en ligne) qui sont devenues la norme ces dernières années, au lieu de cela, nous nous concentrerons sur des aspects de la fonctionnalité.

3.2. Historique

Le développement du premier outil bayésien, au-delà de la mise en œuvre de l'algorithme, s'est produit simultanément à la flambée de la des recherches sur les réseaux bayésiens en 1989. Hugin [MAD 05] Hugin a été initialement développée par un groupe de l'Université d'Aalborg, dans le cadre d'un projet ESPRIT qui a également produit le système MUNIN. Le développement de Hugin s'est poursuivi à travers un autre projet Lauritzen-Jensen appelé ODIN. Hugin Expert a été créé pour commencer à commercialiser l'outil Hugin. Hugin Expert a toujours contribué et profité de la dernière recherche sur les réseaux bayésie. En 1998, Hewlett-Packard a acheté 45% de Hugin Expert.

Le développement du BNJ a débuté en septembre 1997 au département d'informatique de l'Université de l'Illinois. BNJ sont une suite d'outils logiciels open-source pour la recherche et le développement utilisant des modèles graphiques de probabilité. Il est publié par le laboratoire de l'Université du Kansas pour la découverte des connaissances dans les bases de données (KDD).

GeNIe [DRU 09] est un environnement de développement réalisé par DSL (Decision Systems Laboratory) de l'université Pittsburgh en 1998 et conçu pour la construction de modèles décisionnels théoriques et il définit l'interface originale pour SMILE. Il est implémenté dans Visual C ++ et s'appuie fortement sur MFC (Microsoft Foundation Classes).

BayesiaLab est conçu par Bayesia S.A.S. qui est une société française de développement de logiciels fondée en 2001 par le Dr Lionel Jouffe et le Dr Paul Munteanu, spécialisé dans la technologie de l'intelligence artificielle. BayesiaLab est un puissant outil d'intelligence artificielle qui fournit aux scientifiques un environnement complet pour l'apprentissage automatique, la modélisation des connaissances, l'analyse, la simulation et l'optimisation, tous basés sur le paradigme du réseau bayésien.

3.3. Les logiciels bayésiens

Les réseaux bayésiens associent étroitement une structure de graphe (nœud et arc) et une information probabiliste (table de probabilités) en attribuant à chaque nœud du graphe une variable aléatoire. Un réseau bayésien peut être appris à partir de base de données et/ou modélisé par un expert. Il est ensuite possible de mettre à jour les probabilités d'occurrence de chaque état des variables en fonction d'informations sur l'état d'autres variables.

Il y a des logiciels de manipulation graphique de réseaux bayésiens. Ils ont pour but de permettre la saisie, la modification, l'utilisation et l'apprentissage de modèles à base de réseaux bayésiens.

3.3.1. Hugin

La société danoise Hugin Expert A/S [MAD 05], qui édite ce logiciel présenté dans la (Fig.3.1), a été créée en 1989 et est située à Aalborg au Danemark. La société a été créée suite à un projet ESPRIT, qui avait pour but de développer des systèmes experts de diagnostic dans le domaine médical. Hugin est un outil de construction des réseaux bayésiens, probablement le plus connu et le plus utilisé commercialement il offre version limitée téléchargeable pour les systèmes d'exploitation : Linux et Windows.

. Cet outil présente les fonctions principales suivantes :

- Construction de bases de connaissance fondées sur des réseaux bayésiens ou des diagrammes d'influence ;
- Développement de réseaux bayésiens orientés objets ;
- Interface graphique intuitive et simple à utiliser.
- Fournit la fonction d'apprentissage de structure et de paramètres.

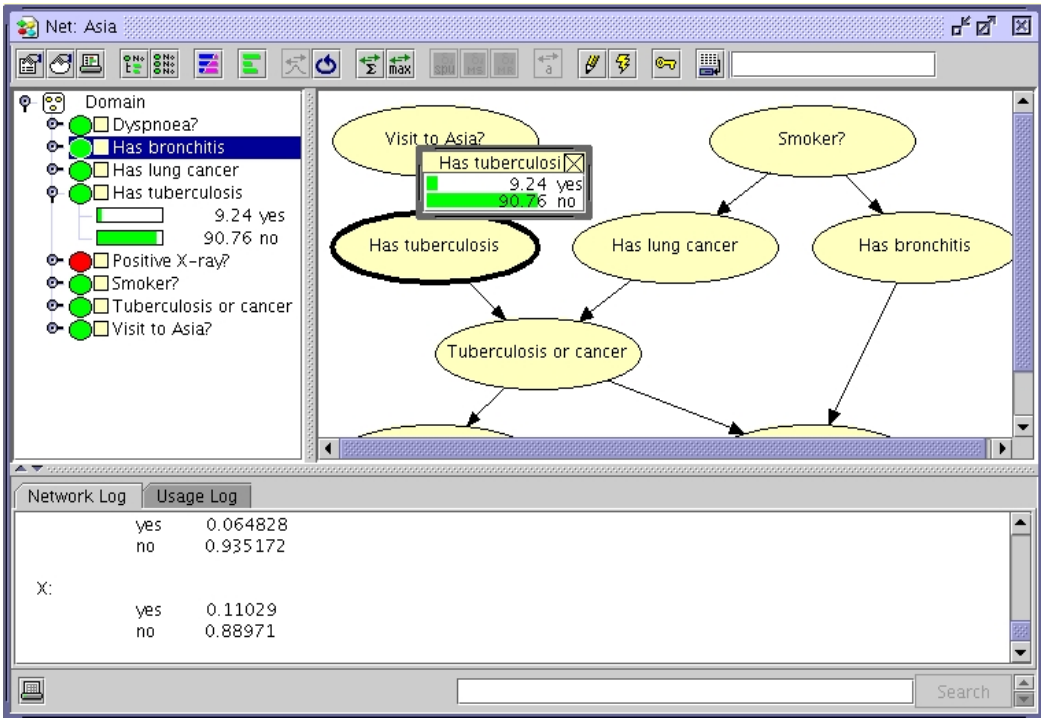


Fig.3. 1 La fenêtre réseau d'un réseau Hugin fonctionnant en mode Run

- **Inférence**

L'algorithme de base est l'algorithme de l'arbre de jonction, avec des options pour choisir entre les variations. L'arbre de jonction peut être consulté. Il y a la possibilité de faire varier la méthode de triangulation. Une version approximative de l'algorithme de l'arbre de jonction est proposée.

Le mode le plus simple d'inférence consiste à entrer des observations dans le réseau, simplement en cliquant sur la valeur observée. Le type d'inférence standard, c'est-à-dire le calcul de la probabilité des nœuds non observés conditionnellement aux observations, s'appelle la propagation Sum normal dans Hugin, qui offre d'autres modes d'inférences. En particulier, la propagation Max normal permet de trouver la configuration du réseau la plus probable, ayant effectué certaines observations.

3.3.2. Bayesian network tools in Java « BNJ »

BNJ (Bayesian network tools in Java) [HSU 02] est un ensemble d'outils Java de recherche et de développement des réseaux bayésiens. Ce projet a été développé au sein du laboratoire KDD « Knowledge Discovery in Databases» de l'université du Kensas. C'est un projet Open source distribué sous la licence GNU (General Public Licence).

Sa dernière version 3.3+ a été publiée en Avril 2006. Cette version fournit une interface graphique comme le montre la (fig.3.2) qui facilite la création, la modification, l'importation et l'exportation des réseaux bayésiens. Elle fournit aussi un ensemble d'algorithmes d'inférence pour les réseaux bayésiens. On détaille dans la suite le contenu de cette boite à outils dans sa version 3.3+. Pour la définition des réseaux bayésiens dans l'environnement de BNJ v3.3+, l'utilisateur peut utiliser l'interface graphique de ce système. Il est possible de définir deux types de distribution de probabilité pour les nœuds : distribution tabulaire discrète et distribution continue. Les réseaux bayésiens crées sont stockés dans des fichiers XML.

- **Inférence**

BNJ v3.3+ fournit un ensemble d'algorithmes d'inférence pour les réseaux bayésiens. Ces algorithmes se classent en deux catégories : inférence exacte et inférence approchée.

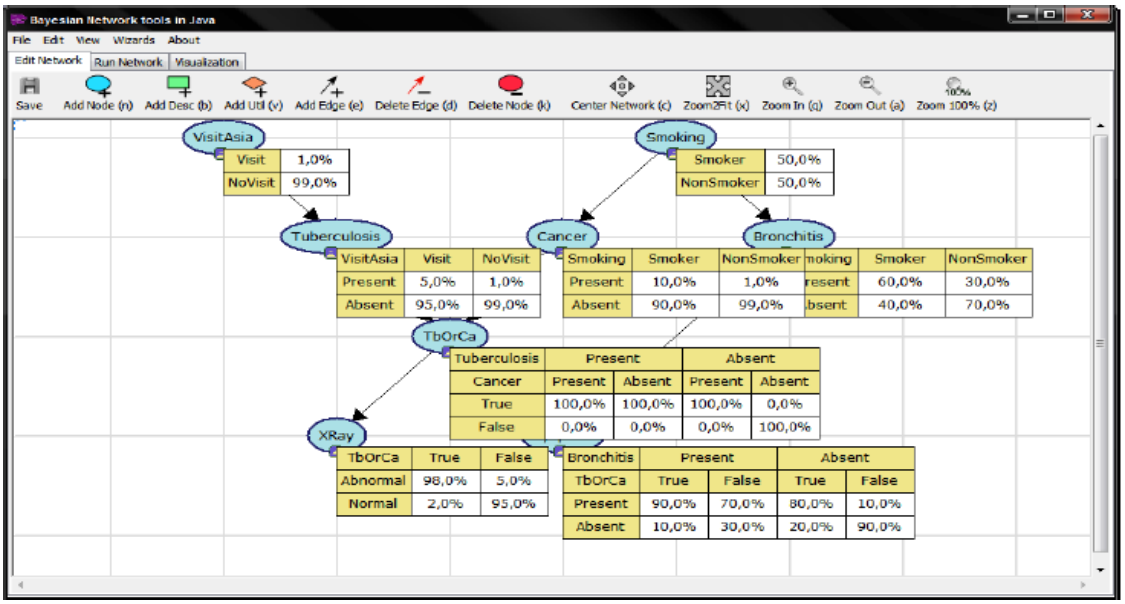


Fig.3. 2 Visualisation de l'exemple de visit-Asia sous BNJ

- **Les algorithmes d'inférence exacte développés sont :**

"Arbre de Jonction", "Elimination des variables avec optimisation", "Singly-connected network belief propagation"("Pearl") et "Cutset Conditioning".

- **Les algorithmes d'inférence approchée développés sont :**
  Certaines méthodes utilisent la notion d'échantillonnage tel que "Adaptive Importance Sampling (AIS)", " Logic Sampling" et "Forward Sampling", d'autres méthodes appliquent les algorithmes d'inférence exacte sur une sélection d'arcs du graphe à traiter tels que "KruskalPolytree", "BCS" et "PTReduction".

L'ensemble des fichiers sources de la boite à outils ne comporte pas d'implémentation des algorithmes d'apprentissage de paramètres ni de structure.

La librairie BNJ est structurée de façon arborescente facilitant ainsi la navigation dans les différents répertoires. Les fichiers du code source sont bien soignés et présentent des informations utiles pour la compréhension du rôle de chaque fichier.

En effet, les différentes méthodes sont décrites par un commentaire, de plus, les variables et les méthodes portent des noms significatifs. Le site de BNJ a deux fichiers documentant la version BNJ 2.03 a, un destiné aux nouveaux utilisateurs BNJ et un autre destiné aux développeurs qui s'intéressent au code source. Mais la dernière version est fournie sans aucune documentation.

Le soin de la structure et du contenu des fichiers sources s'avère alors le seul refuge des nouveaux développeurs.

3.3.3. GeNIe

GeNIe constitue l'interface graphique de cette bibliothèque implémentée dans Visual C ++. SMILE et Genie sont développés par DSL (Decision Systems Laboratory) de l'université Pittsburgh. Ils sont libres et téléchargeables sur le site http://genie.sis.pitt.edu/

SMILE (Structural Modeling, Inference, and Learning Engine) est une bibliothèque de GeNIe entièrement portable de classes C++. Il permet l'implémentation des méthodes basées sur la théorie de la décision, les réseaux bayésiens et les diagrammes d'influence ; nommé aussi un moteur d'inférence de réseau bayésien.

Pour accéder à la bibliothèque SMILE à partir d'autres langages de programmation, certains "wrappers" sont développés : jSMILE pour Java, SMILE.NET pour un environnement Microsoft .NET et pocketSMILE pour Pocket PC. Complémentaire à la plate-forme SMILE et à ses wrappers, le développement de SmileX, un composant Windows ActiveX qui permet d'accéder à SMILE à partir de n'importe quel environnement de programmation Windows, y compris les pages Web [THI 99].

GeNIe permettent de créer, éditer, enregistrer et charger des modèles graphiques telles que : les diagrammes d'influences, les réseaux bayésiens dynamiques et les utiliser pour un raisonnement probabiliste. Cet outil supporte l'apprentissage de la structure et des paramètres et les algorithmes d'inférence multiples. En plus de cela, il existe quelques caractéristiques spéciales pour le diagnostic comme la fourniture d'une interface montré dans la (Fig.3.3) pour effectuer un diagnostic. Il offre une possibilité de stocker des fichiers de format XML.

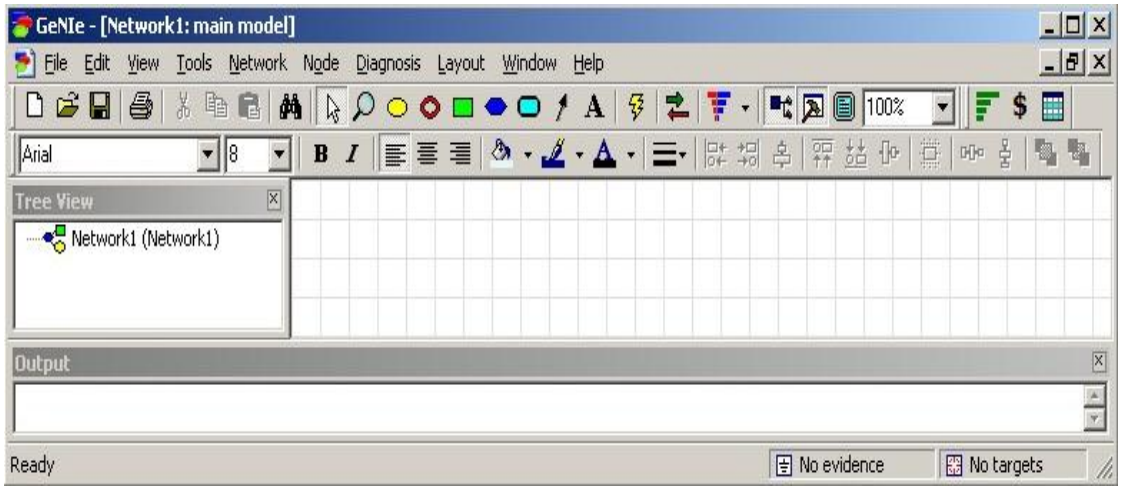


Fig.3. 3 interface principale de GeNIe

- **Inférence**

GeNIe fournit un ensemble d'algorithmes d'inférence pour les réseaux bayésiens dans les deux catégories : inférence exacte et inférence approchée. La plupart des algorithmes de GeNie sont simple. Sur les algorithmes disponibles deux peuvent être considérées comme destinées à inférence rapide réelle : Clustering et Epis-échantillonnage

Clustering, aussi connu comme la *propagation de l'arbre de clique* ou l'algorithme *d'arbre de jonction*, Est le principal algorithme de solution exacte dans GeNIe. Il est relativement rapide, et il peut être utilisé pour faire l'inférence même dans les réseaux

modérément grands. Cependant, le cas d'un réseau très grand nombre de nœuds de hasard, les algorithmes approximatifs sont généralement plus pratiques.

Le code source de GeNIe and SMILE est fermé Contrairement à SMILE.NET et jSMILE sont de l'open source. La paire offre une combinaison de prise en charge de l'interface graphique et interfaces de langage de programmation. La plupart des fonctionnalités souhaitées sont présentes, avec Les seuls revers majeurs étant le code source fermé, et le soutien limité et mal documentées pour les nœuds non discrets. GeNIe prend la tête avec toutes ces caractéristiques en comparant avec les logiciels cités précédemment.

3.3.4. BayesiaLab

BayesiaLab est un laboratoire complet de manipulation de réseaux bayésiens qui permet d'élaborer des modèles décisionnels par recueil d'expertise et automatiquement à partir des données, d'assimiler rapidement des connaissances représentées grâce à une boite à outils d'analyse originale, d'exploiter des modèles en mode interactif ou par lots et faire l'apprentissage des politiques d'actions. On peut consulter aussi le site htp:/www.bayesia.com/ pour avoir plus d'information [NAÏ 07].

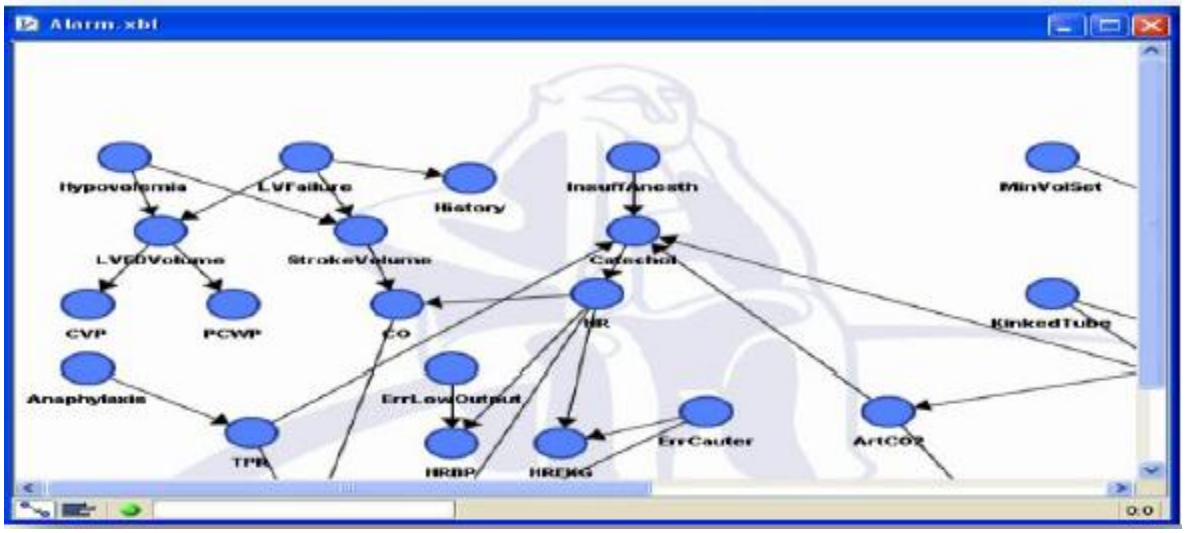


Fig.3. 4 Visualisation des graphes dans bayesiaLab

Dans le domaine de data mining, les méthodes d'Intelligence Artificielle peuvent résoudre à plusieurs types de problèmes comme la classification pour prédire les valeurs catégorielles des variables, la régression pour prédire les valeurs numériques des variables,

etc. En, général on les divise en deux groupes : méthodes explicatives et méthodes non-explicatives.

Le changement des paramètres va influencer la taille de la mémoire utilisée, le temps d'exécution des algorithmes et le résultat obtenu. Pour avoir de bons résultats avec des ressources matérielles limitées, il est nécessaire de choisir les valeurs appropriées des paramètres.

- **Domaines d'application (BayesiaLab)**
  - Modélisation de systèmes complexes (processus industriels).
  - Analyse globale de risque et politique de sécurité (réseau de transport ferroviaire).
  - Marketing (laboration d'un profil client face un produit cible).
  - Risk manager.
  - Data Mining des bases clients (marketing et gestion des fraudes).
  - Détection des intrusions.
- **Inférence**

Le logiciel gère deux types d'inférence : exacte (basée sur l'algorithme de l'arbre de jonction) et une inférence approchée lorsque les réseaux sont de complexité trop grande. L'approximation peut se faire soit par échantillonnage stochastique (Likelihood Weighting), soit par inférence exacte sur un graphe simplifié (suppression des relations les plus faibles et causant la plus grande complexité). Pour les réseaux de grande taille, un mode d'inférence exacte bas sur les requêtes est également disponible (relevance reasoning). Ce mode permet, par l'analyse des observations et des nœuds requêtes, de construire l'arbre de jonction minimal. L'exploitation nécessite la possibilité d'insérer des observations dans le réseau. BayesiaLab permet d'insérer des évidences certaines positives ou négatives (ce nœud a cette valeur ou n'a pas cette valeur), des vraisemblances (une valeur entre 0 et 100 sur chaque modalité), et des distributions de probabilités.

La console est le lieu des messages non préemptifs de BayesiaLab à l'utilisateur. Elle permet de visualiser (par exemple) des valeurs comme représenté dans la (Fig.3.5) précises lors d'apprentissage, d'inférence, etc.

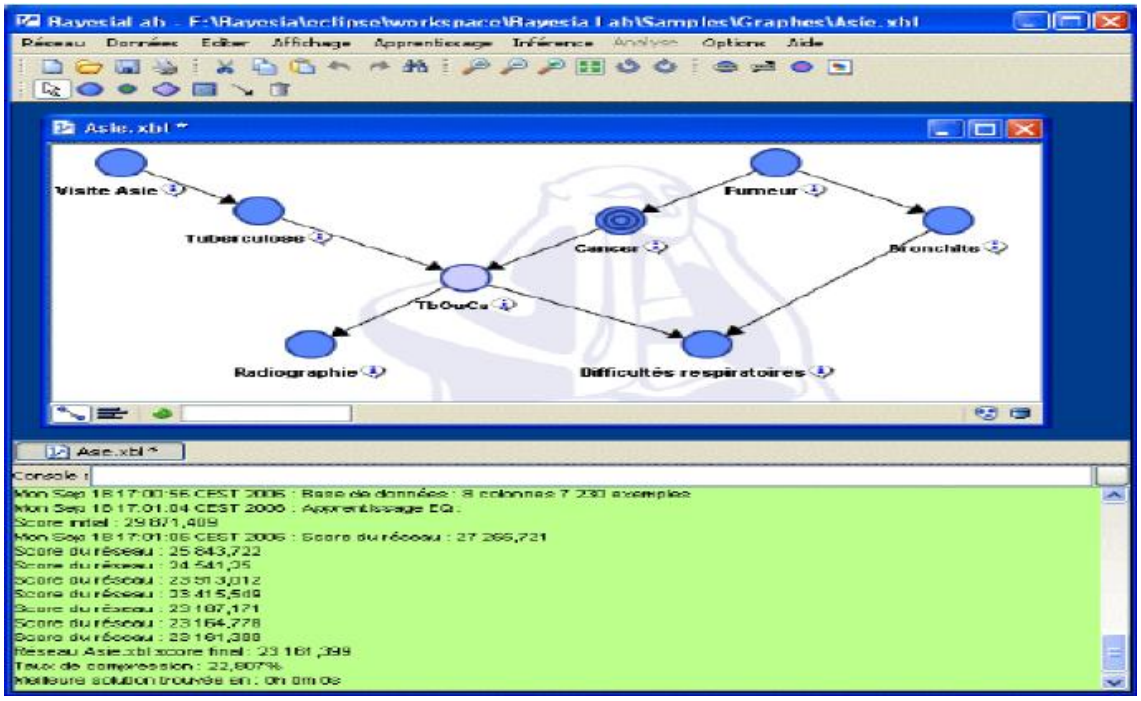


Fig.3. 5 Affichage des résultats dans la console de BayesiaLab

BayesiaLab exploite le réseau bayésien en interactif (à partir d'observations entrées manuellement des moniteurs ou automatiquement par un fichier d'observations) ou en « batch ».

- **Apprentissage**

L'apprentissage est un des points forts de BayesiaLab. Il utilise des méthodes et des algorithmes qui sont à la pointe de la recherche dans le domaine (les fondateurs de BayesiaLab étant des chercheurs spécialisés dans l'apprentissage et particulièrement dans

L'apprentissage dans BayesiaLab prend comme entrée un fichier texte ou un lien ODBC4 décrivant l'ensemble des cas (un cas par ligne ou un cas par colonne). Ce fichier peut intégrer un ensemble de caractères indiquant les valeurs manquantes.

Les assistants d'importation permettent la configuration de la lecture (séparateurs, ligne de titre, valeurs manquantes, transposition), l'échantillonnage, la sélection des colonnes importer, le typage de ces colonnes (variable discrète ou continue, variable de pondération des individus, individu d'apprentissage ou de test), la scission de la base en ensembles d'apprentissage et de test.

En tant que laboratoire d'étude de réseaux bayésiens, BayesiaLab offre un trés large choix dans les algorithmes à utiliser pour exploiter ces données. Il propose :

- La prise en compte de la connaissance experte exprimée sous la forme d'un graphe initial et d'un nombre de cas équivalents, des indices temporels des variables (pas d'ajout d'arc entre le futur vers le passé), des contraintes définies sur les nœuds et les classes. L'apprentissage de réseaux bayésiens).
- Une gestion rigoureuse des valeurs manquantes.
- Une fonction de stratification, ainsi que la prise en compte d'une variable de pondération (coefficient de redressement).
- Une complexité structurelle modifiable (jouant le rôle de seuil de significativité).
- Un apprentissage des paramètres (tables de probabilités).
- La découverte d'associations pour mettre en évidence l'ensemble des relations probabilistes directes présentes dans les données.
- La recherche commence généralement par un graphe non connecté, mais il est également possible de commencer à partir d'une structure initiale (fournie par un expert ou résultant d'un précédent apprentissage). Sauf s'ils sont fixés par l'expert, les arcs pourront alors être remis en cause lors de l'apprentissage.

➢ **Cinq algorithmes sont proposés (BayesiaLab) :**

- Arbre de recouvrement maximal,
- Deux algorithmes de recherche dans les classes d'équivalence,
- Une recherche Taboo dans l'espace des RB,
- Une recherche Taboo dans l'espace des ordres de nœuds.
- Caractérisation probabiliste d'un nœud cible (apprentissage entièrement focalisé sur ce nœud cible).

3.4. Etude comparative entre les différents logiciels bayésiens

Le (Tableau.3.1) ci-dessous couvre les informations de base pour les trois logiciels, ainsi d'autre logiciels telles que les informations techniques sur la disponibilité des sources, les plates-formes, l'interface graphique, les types de nœuds pris en charge (c'est-à-dire discrets et / ou continus), le logiciel permet l'apprentissage ou non (paramètres et / ou structure), décrit les principaux algorithmes d'inférence et indique si le logiciel est gratuit ou commercial [CHR 09].

Nom	*SRC*	*API*	*EXEC*	*IUG*	*D/C*	*Param*	*Struc*	*Infer*	*Free*
baysiaLab	N	N	-	O	Cd	O	O	JT, G	$
GeNie	C++	O	W, U	O	D	O	O	JT (+S)	O
BNJ	J	O	-	O	D	N	O	JT, IS	O
Hugin	Y	N	W, U	O	G	O	O	JT	$
BayesBuilder	N	N	W	O	Cd	O	O	JT,G	$
CABeN	C	O	W, U	N	D	N	N	S(++)	O
CaMML	C	N	U	N	Cx	O	O	N	O
Ergo	N	O	W, M	O	D	N	N	JT(+S)	$
GMTK	N	O	U	N	D	N	N	MC	O
Java Bayes	J	O	W, U, M	O	D	N	N	JT	$
PNL	C++	-	-	N	D	O	O	JT	O
UnBBayes	J	O	-	O	D	N	O	JT	O

Tableau.3. 1 La comparaison des fonctionnalités des logiciels

- **SRC :** Le code source est-il inclus ?

  N = non.

  Si oui, quelle langue ? J = Java, M = Matlab, L = Lisp, C, C ++.
- **API :** Une interface de programme d'application est-elle incluse ?

  N= signifie que le programme ne peut pas être intégré dans votre code, c'est-à-dire qu'il doit être exécuté comme un exécutable autonome. O=signifie qu'il peut être intégré.
- **EXEC :** L'exécutable s'exécute sur : W = Wfffffffffindows (95/98/2000 / NT), U = Unix, M = Mac- Intosh, - = Toute machine avec un compilateur.
- **IUG :** Une interface utilisateur graphique est-elle incluse ? O = Oui, N = Non.
- **D/C :** Les nœuds à valeur continue sont-ils pris en charge (aussi bien que discrets) ?
  - G = (conditionnellement) nœuds de Gaussians pris en charge analytiquement,
  - Cs = nœuds continus pris en charge par l'échantillonnage,
  - Cd = nœuds continus pris en charge par la discrétisation,
  - Cx = nœuds continus supportés par une méthode non spécifiée,
  - D = seuls les nœuds discrets sont pris en charge.

- **Params :** La fonctionnalité du logiciel comprend-elle l'apprentissage des paramètres ? O = Oui, N = Non.
- **Struc :** La fonctionnalité du logiciel comprend-elle l'apprentissage de la structure ? O = Oui, N = Non.
- **Infer :** Quel algorithme d'inférence est utilisé ?
  - JT = Arbre de jonction,
  - S = Echantillonnage,
  - IS = Importance échantillonnage,
  - GS = Echantillonnage de Gibbs
  - ++ = De nombreuses méthodes ont été fournies,
  - + S= Certains paquets supportent une forme d'échantillonnage (Pondération de vraisemblance, MDMC), en plus de leur algorithme exact.
  - ? = Non spécifié,
- **Free :** Une version gratuite est-elle disponible ?
  - O = Gratuit (mais peut-être uniquement pour une utilisation académique)
  - $ = Commercial (bien que la plupart possèdent des versions gratuites qui sont restreintes de diverses façons, par exemple, la taille du modèle est limitée ou les modèles ne peuvent pas être sauvegardés).

3.5. Domaines d'application des réseaux Bayésiens

- **Santé**

Les premières applications des réseaux bayésiens ont été développées dans le domaine du diagnostic médical. Les réseaux bayésiens sont particulièrement adaptés à ce domaine parce qu'ils offrent la possibilité d'intégrer des sources de connaissances hétérogènes (expertise humaine et données statistiques), et surtout parce que leur capacité à traiter des requêtes complexes (explication la plus probable, action la plus appropriée) peuvent constituer une aide véritable et interactive pour le praticien.

Le système Pathfinder, développé au début des années 1990 a été conçu pour fournir une assistance au diagnostic histopathologique, c'est-à-dire basé sur l'analyse des biopsies. Il est aujourd'hui intégré au produit Intellipath, qui couvre un domaine d'une trentaine de types de pathologies [BEC 99] [HAL 04].

- **Industrie**

Dans le domaine industriel, les réseaux bayésiens présentent également certains avantages par rapport aux autres techniques d'intelligence artificielle par exemple la société danoise

Hugin, considérée comme l'un des pionniers dans le développement des réseaux bayésiens. Hugin a développé pour le compte de Lockheed Martin le système de contrôle d'un véhicule sous-marin autonome. Ce système évalue en permanence les capacités du véhicule à réagir à certains types d'événements.

- **Défense**

La fusion de données est particulièrement un domaine d'application privilège des réseaux bayésiens, grâce à leur capacité à prendre en compte des données incomplètes ou incertaines et guider la recherche ou la vérification de ces informations. La société Mitre a développé un système de défense tactique embarqué pour les navires de guerre de la marine amé décide des ripostes à adopter. Ce système analyse les informations permet en particulier de gérer les menaces multiples, qui peuvent générer des conflits sur l'affectation des armes.

- **Banque/finance**

Les applications dans le domaine de la banque et de la finance sont encore rares, ou du moins ne sont pas publiées. Mais cette technologie présente un potentiel très important pour un certain nombre d'applications relevant de ce domaine, comme l'analyse financière, le scoring, l'évaluation du risque ou la détection de fraudes.

- **L'informatique**

L'utilisation de réseaux bayésiens dans les agents bureautiques a été largement développée par Microsoft dans les logiciels d'aide et de diagnostic pour son système d'exploitation Windows, partir de Windows 98. De m me, l'agent Office Assistant est un système d'aide proactif intégré dans Office, à partir de la version 97. Plusieurs agents de support technique de Microsoft ont également été développés dans le cadre du projet LUMIERE du groupe DTAS (Decision Theory and Adaptive Systems).

L'application Vista, peut également très considérée comme un agent intelligent, dont le rôle est de sélectionner les données présentées à un utilisateur en fonction de l'état du système physique qu'il doit superviser.

Les réseaux bayésiens constituent le modèle idéal pour embarquer de l'intelligence ou de la connaissance. Embarquer de l'intelligence revient doter un agent d'un équipement lui permettant de décider dans des environnements incertains, et de s'adapter lorsque ces environnements changent. Un module bayésien de prise de décision, éventuellement capable d'adaptation, est l'un des meilleurs équipements que l'on puisse fournir un agent envoyé en

mission sur Internet, ou sur d'autres types de réseau, l'information est par nature incertaine et évolutive, voire manipulée.

3.6. L'incertitude

Le cœur du problème de décision en général et de décision médicale en particulier, réside dans la nature de l'incertitude à laquelle le décideur est confronté et de la façon dont celui-ci l'appréhende.

La santé est un domaine où l'incertitude prend une importance considérable. En effet, la morbidité n'est pas indépendante des facteurs de risque. Aussi, le traitement de ce risque a des répercussions fondamentales sur la pratique médicale. C'est pourquoi la décision médicale en contexte d'incertitude a été appréhendée par trois types d'approches [BEN 10]:

- La théorie statistique de l'information qui quantifie le degré d'incertitude par la probabilité de présence d'un événement (la maladie par exemple). Plus l'incertitude est grande, plus l'information qui la quantifie est élevée. Dans le domaine médical, une probabilité de présence de la maladie de 0 ou de 1 (maladie respectivement absente ou certaine) conduit à une information nulle.
- La théorie statistique de révision bayésienne des croyances pour la quelle patient et médecin doivent prendre leur décision en contexte d'incomplétude et d'asymétrie d'information. L'expertise médicale se réduit à la transformation des croyances issues des évaluations probabilistes en certitudes (ou quasi-certitudes) grâce à l'appréciation des informations fournies par le patient.
- La théorie économique de l'information, représentée par l'axiomatique de l'utilité espérée (Von Neumann et Morgenstern, 1944) qui formalise l'intuition ancienne de Bernoulli selon laquelle ce que maximisent les individus n'est pas l'espérance mathématique des gains mais celle de l'utilité de ces gains.

3.6.1. L'incertitude médicale

La décision médicale est toujours associée à un degré d'incertitude qui est inhérent au domaine de la biologie et de l'humain. Ceci est dû à la difficulté, voire l'impossibilité de recueillir certaines données physiopathologiques, de faire des mesures et examens sans déroger à la déontologie, sans parler de la variabilité des sujets ou de la rareté de certaines ressources.

L'ensemble des données cliniques ou para cliniques est également difficile à synthétiser et il est donc fréquent de tirer des conclusions incertaines qui amèneront des décisions

diagnostiques et thérapeutiques dont on sait qu'elles sont susceptibles d'être remises en question.

C'est ce qui explique d'une part le développement des méthodes probabilistes non explicatives et d'autre part le développement des systèmes experts qui essayaient de simuler les raisonnements des experts et se voulaient plus transparents en permettant d'objectiver les connaissances des experts et le raisonnement suivi.

Il n'en reste pas moins que ce raisonnement des experts est basé sur des inférences prenant en compte l'incertain tant au niveau individuel des patients qu'au niveau de la population. Les recherches actuelles tendent donc à combiner les deux approches en utilisant des raisonnements basés sur les réseaux probabilistes introduits par Pearl sous le terme de 'Belief Networks'[LEB 94].

3.6.2. Réseau bayésien et l'incertitude médicale

L'action médicale repose sur la capacité de raisonnement du médecin et son aptitude à prendre des décisions alors que les informations utilisées sont potentiellement entachées d'incertitude. Cette incertitude est d'origine multiple : possibilité d'erreur dans les données ambiguïté de la représentation de l'information, incertitude sur les relations entre les diverses informations. Une première approche à la représentation de la connaissance dans le contexte d'incertitude a utilisé la théorie des probabilités. Ainsi, plusieurs études ont montré que les systèmes d'aide à la décision élaborés sous le modèle probabiliste pouvaient faire aussi bien, voire même mieux que le médecin [PEW 01].

Cependant, bien qu'utile pour codifier l'incertitude dans un problème de décision, l'approche purement probabiliste comporte des limites. Tout d'abord, l'élaboration, la représentation et la manipulation de telles bases de connaissances sont souvent nécessaire de postuler des hypothèses insuffisantes pour obtenir un réel impact sur la pratique médicale.

Enfin les alternatives à une décision ou les préférences de l'utilisation ne peuvent être prises en compte. Les systèmes basés sur des règles, apparus dans les années 1970, circonvenaient la difficulté du calcul probabiliste en utilisant d'autres paramètres pour représenter l'incertitude : facteur de certitude, calcul de Dempster-Shafer ou logique floue. Plus récemment, le développement de la technologie informatique a permis de reconsidérer le formalisme probabiliste. Ceci est particulièrement le cas du domaine de l'analyse de décision, méthodologie basée sur la théorie des probabilités mais qui permet de représenter explicitement les problèmes de décision et les préférences de l'utilisateur. Les réseaux

bayésiens ou diagramme d'influence (Belief network) constituent un des modèles de représentation des connaissances utilisables en analyse de décision [PEW 01].

Un réseau bayésien, également appelé diagramme d'influence ou réseau causal est un graphe dirigé acyclique dans lequel les nœuds représentent les variables et les arcs précisent les dépendances probabilistes entre les variables. Il permet d'afficher graphiquement les variables d'un problème de décision et les relations ou influences entre variables, qui peuvent mener à des décisions complexes. Les réseaux bayésiens centrent l'attention de la décision exclusivement sur les composants du problème en relation avec la tache de décision présente. Exclure l'information non pertinente du diagramme d'influence facilite le travail de la décision et lui permet de gagne du temps puisqu'il existe moins de variables à interpréter [HEN 92].

3.7. Conclusion

Les réseaux bayésiens représentent une approche de choix dans la représentation de connaissances incertaines et dans l'exploitation de celles-ci. Par ailleurs, plusieurs domaines sont intéressés par ce type de représentation, de ce fait de nombreux logiciels existent pour saisir et traiter des réseaux bayésiens. Ces logiciels présentent des fonctionnalités plus ou moins évoluées : apprentissage des probabilités, apprentissage de la structure de réseau bayésien, possibilité d'intégrer des variables continues, des variables d'utilité de décision [NAï 07] etc.

Le chapitre suivant nous allons aborder l'aspect conception du système SADMsein (Système Aide au Diagnostic des Maladies des Seins), et nous avons combiné et réalisé les idées et les concepts développés dans les chapitres précédents ainsi l'outil de développement utilisé dans ce contexte.

Chapitre 4 Conception du système SADMseins

4.1. Introduction

Le but de ce travail est la construction d'un système d'aide au diagnostic médical qui nécessite presque toujours la prise en compte de l'incertitude dans le raisonnement, ce qui explique le développement des méthodes probabilistes. Les réseaux bayésiens sont des outils privilégiés pour les problèmes de diagnostic. Au cours de cette dernière décennie, ils ont été utilisés avec succès dans le domaine médical et font l'objet, actuellement des recherches intenses avec différents objectifs et d''évaluation des connaissances.

Dans ce travail nous modélisons un système expert qui essaie de simuler le raisonnement des médecins pour l'aide au diagnostic des pathologies les plus fréquentes des seins. La structure du réseau bayésien est évaluée uniquement à partir des avis des experts, elle permet une représentation de connaissances qualitatives et quantitatives exprimant l'incertitude décomposée en quatre niveaux : niveau clinique, niveau biologique, niveau imagerie médicale et niveau diagnostic.

Une fois la structure de réseau est mise en place et que les tables de probabilités sont définies, les réseaux bayésiens sont utilisés pour calculer les probabilités. Ce procède est appelé l'inférence bayésienne.

L'inférence bayésienne est présentée par Naim [NAI 07] comme « le processus de propager une ou plusieurs informations certaines au sein d'un réseau pour en déduire comment sont modifiées les croyances concernant les autres nœuds. En d'autres termes, l'inférence sert à calculer la probabilité d'une hypothèse suite à l'observation des évidences.
Les évidences correspondent aux nœuds d'entrée et les hypothèses sont les différents états des nœuds de sortie du réseau. L'injection des probabilités des nœuds d'entrée va modifier récursivement les probabilités des nœuds enfants jusqu'aux nœuds de sortie.

Nous présentons dans un second temps, l'architecture générale et les différents modules qui composent notre système « outil GeNIe ». Enfin nous montrons le principe détaillé de l'algorithme choisi et son déroulement.

4.2. Description générale de processus du diagnostic

Dans le cas de diagnostic des maladies, plusieurs symptômes identiques peuvent aboutir à des maladies différentes. Pour ce faire, nous proposons ci-dessous (Fig.4.1) un système d'aide au diagnostic des pathologies du sein. Le raisonnement médical part d'un

examen clinique cognitif qui est un ensemble de données interrogatoires. En effet, à partir des symptômes observés (les signes cliniques), le médecin identifie les facteurs prédisposants et sur cette base plusieurs hypothèses sont alors établies. Pour confirmer ses hypothèses, le médecin doit faire passer la patiente par un examen radiologique (mammographie). On sait qu'un examen clinique n'est pas certain à 100%. Il reste toujours une incertitude, c'est qu'à travers l'examen biologique (biopsie, cytoponction) que le médecin arrive à poser son diagnostic final. La **(**Fig.4.1) présente les différentes étapes du diagnostic [NAO 07] [DJE 10].

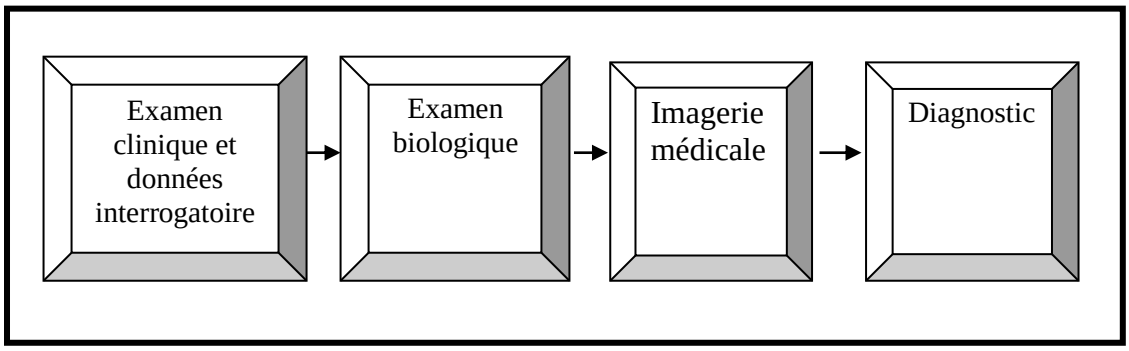


Fig.4. 1 Description du problème du diagnostic.

4.3. La conception du SADMsein

L'utilisation du formalisme des réseaux bayésiens nous aide à présenter les informations pertinentes pour un traitement efficace de notre problème. Suites aux décisions prises avec les experts du domaine, ces informations (variables) sont réparties en quatre niveaux sachant qu'elles sont discrètes.

4.3.1. Définition des variables du réseau

- Niveau clinique : constitué d'un ensemble de facteurs permettant la détection de différentes maladies du sein exemple :
    - Age.
    - Sexe.
    - Antécédents familiaux.
    - Mastodynie.
    - Rétraction du mamelon.
    - Modification de la coloration des téguments.
    - Ecoulement du mamelon.
    - Adénopathie.

- Masse dans le sien.

- Niveau imagerie médicale : comprend le résultat de l'image mammographie du sein qui indique la présence d'une anomalie ou pas sous forme d'une masse accompagné par un rapport descriptif de la masse (le résultat de dépistage de la mammographie) citons : la taille de la masse, la forme, le contour, l'homogénéité, micro calcification.
- Niveau biologique : présente l'état d'analyse médicale qui est le résultat de la biopsie et la cytoponction, ces deux critères ont une très grande influence dans le résultat de diagnostic des maladies des seins.
- Niveau diagnostic : représente le diagnostic final avec thérapie pour un ensemble de maladies les plus fréquentes du sein : Kyste, Adénome fibrome, lipome, abcès, cancer du sein. Ces informations sont représentées par des variables discrètes.

4.3.2. Construction du réseau bayésien

Nous donnons l'exemple d'un réseau bayésien, (Fig.4.2) qui modélise de manière raisonnable le processus de cinq maladies les plus fréquentes du sein (Kyste, Adénome fibrome, lipome, abcès, cancer du sein). Le mode de raisonnement est ici représenté par un diagramme de causalité (RB).

Dans notre cas la structure du réseau est décrite à l'aide des médecins et les probabilités sont obtenues à partir de l'apprentissage bayésien. Ce processus a conduit à un modèle avec 50 liens et 26 nœuds comme indiqué dans (Fig.4.2) qui présente une modélisation du réseau bayésien de façon claire le processus de diagnostic de cinq maladies du sein validées par un expert (médecin).

Nous avons choisi pour cette étude un outil de réseau bayésien disponible à notre niveau. Il s'agit de GeNie qui est un environnement de développement pour construire des modèles de décision graphiques. Il a été développé à la décision Systems Laboratory, Université de Pittsburgh [DRU 99]. Il dispose d'une bibliothèque appelée SMILE (Structural Modeling, inférence, et l'apprentissage moteur) de fonctions de mise en œuvre des modèles probabilistes et de la théorique de décision graphiques, tels que les réseaux bayésiens. Ses fonctions individuelles, définies dans SMILE Interface Applications de programmeur, permettent de créer, d'éditer, de sauvegarder et de charger des modèles graphiques, et les utiliser pour le raisonnement probabiliste et la prise de décision en situation d'incertitude.

Avec l'outil GeNie Il y a deux façons de construire un réseau bayésien : une construction manuelle ou une construction automatique (dite « d'apprentissage structurelle ») à partir de bases de données. Pour notre cas la construction est manuelle. La première étape

consiste à construire un graphe acyclique orienté, suivi par la deuxième étape afin d'évaluer la distribution de probabilité conditionnelle dans chaque nœud. La première étape dans la conception manuelle d'un réseau probabiliste est d'inclure tous les nœuds (variables aléatoires) à l'aide des éléments fonctionnels importants et une boite à outils qui facilite la construction du réseau bayésien. Les variables, en générale, sont représentées par des cercles (icône ovale) sachant que l'état de nœud peut être identifié : nœuds cibles (de sortie), les nœuds d'observation et les nœuds intermédiaires.

Dans l'évaluation RB nous allons nous concentrer sur les cibles et les nœuds d'observation, les nœuds cibles représentent les états de diagnostic (maladies) qui apparaissent avec des nœuds jaunes, les nœuds d'observation représentent le résultat des observations (symptômes, ou tests) qui apparaissent avec des nœuds bleus, tous ces types sont présentés dans le (Tableau4.1). La démarche suivante consiste à connecter les nœuds en utilisant l'icône de la flèche de la barre d'outils pour définir la dépendance probabiliste entre plusieurs paires de nœuds.

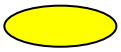
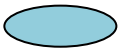
Type de diagnostic et composants du réseau	**Usage typique**
Target 	Pour représenter les maladies
Observation 	Les messages d'erreur, des symptômes ou des tests
Flèche	Représente les arcs qui relient les nœuds parent et les nœuds fils

Tableau.4. 1 Description sur les états des nœuds dans GeNie.

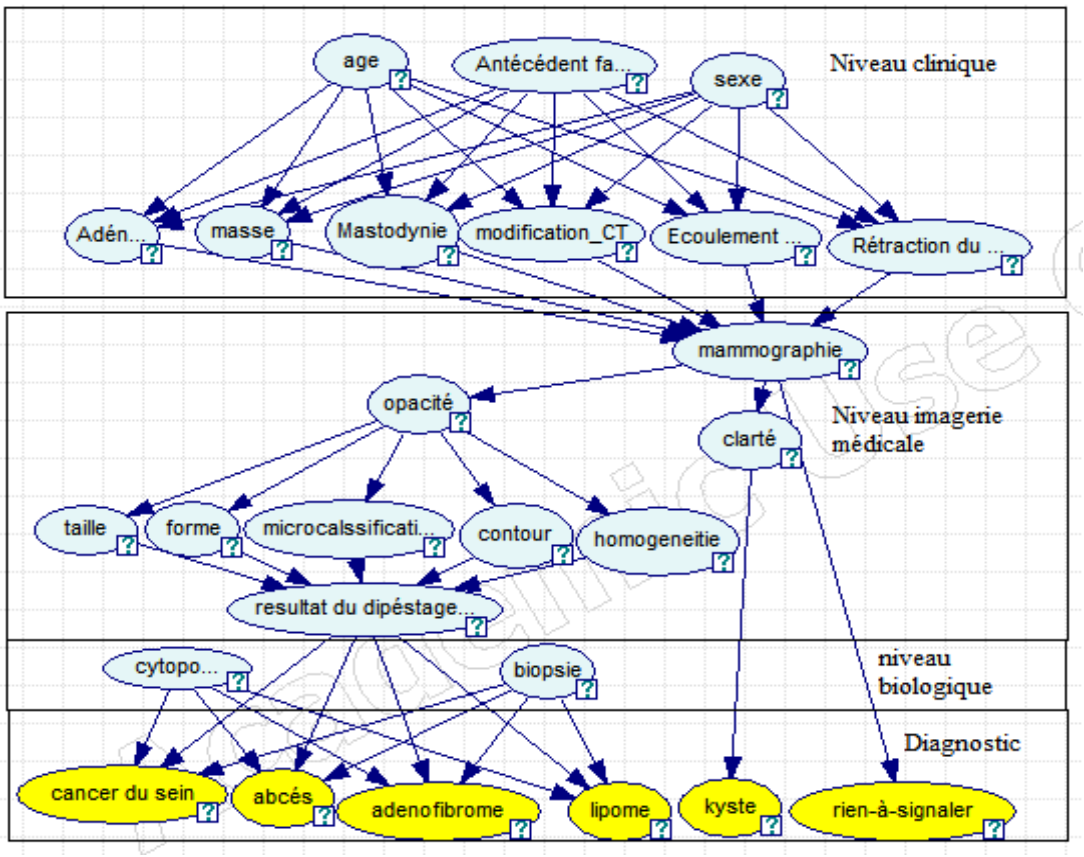


Fig.4. 2 Exemple d'un réseau bayésien avec cinq maladies du sein.

Ce modèle permet de présenter les informations pertinentes pour un traitement efficace de notre problème. Le réseau proposé permet une représentation qualitative et quantitative des connaissances exprimant l'incertitude du diagnostic des maladies les plus fréquentes du sein.

Le graphe orienté acyclique construit, doit inclure des distributions de probabilités conditionnelles pour chaque nœud dans le graphe [LUC 04]. Si les variables sont discrètes, ce qui peut être représenté sous la forme d'une table de distribution (multinomiale), qui indique la probabilité que le nœud enfant prend sur chacune de ses différentes valeurs pour chaque combinaison de valeurs de ses parents. Si la distribution des probabilités conditionnelles ne sont pas disponibles, d'autres méthodes statistiques peuvent être appliquées pour obtenir cette distribution conditionnelle à partir des données (distribution de probabilité conditionnelle par exemple empirique / fréquences estimation). Des méthodes de calcul possibles sont décrites dans [SPI 98].

En premier lieu, nous nous sommes concentrés sur la première instance pour recueillir le maximum d'enregistrements (information) de l'hôpital Ibn Rochd de ANNABA avec l'aide Professeur en gynécologie, Hayet Aures

Nous avons réussi à obtenir 100 cas, les données manquantes ont été estimées par notre médecin, une illustration d'une partie de la base de données est présentée dans la (Fig.4.3). Chaque ligne représente une observation de l'ensemble de données déterminé par des variables représentant les principaux signes, symptômes, tests, les examens et les maladies du sein.

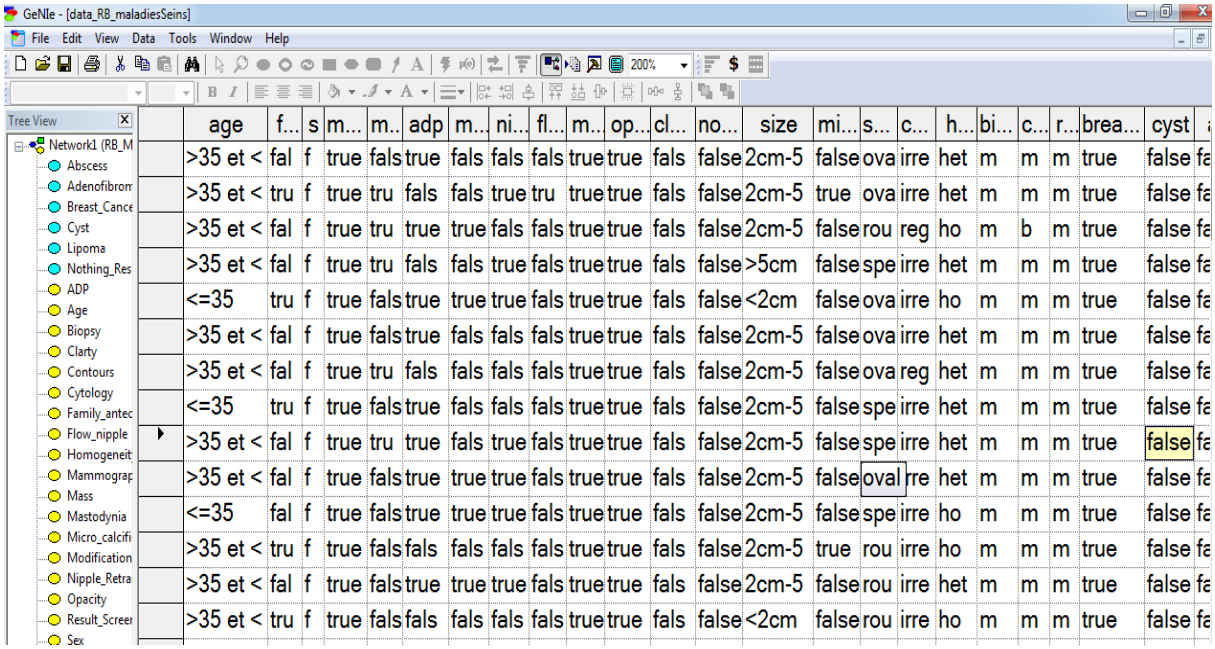


Fig.4. 3 Table des cas retenus pour l'évaluation.

4.3.3. Préparation des données dans le cadre de l'apprentissage automatique

Les réseaux bayésiens utilisent une distribution de probabilité conditionnelle à chaque nœud dans le graphe. Si la distribution de probabilité conditionnelle n'est pas connue, elle peut être obtenue à partir des données par l'estimation de la distribution de probabilité conditionnelle empirique (fréquences conditionnelles).

Dans le cas de l'apprentissage automatique, toutes les variables pertinentes doivent être organisées dans une structure de base de données unique.

Le logiciel GeNie peut apprendre les réseaux bayésiens d'un .dat, .txt, .csv ou fichier ODBC. Si la base de données est dans un format différent (par exemple Microsoft Access ou SAS), le logiciel correspondant par défaut peut généralement traduire le fichier de données dans l'un des formats lisibles.

Pour notre cas nous avons accédé aux données à partir d'une base de données Microsoft Access et importé ce dernier dans GENIE montré la (Fig.4.3) pour apprendre les paramètres

du réseau. Une fois que le fichier est chargé, des données statistiques tirées par la base importée, seront utilisée pour calculer les probabilités conditionnelles en utilisant des algorithmes expérimentaux souvent intégrées dans le logiciel approprié (GENIE) [HOR 14]. Apres l'importation de la base de données il y aura une commande « learn » qui permet la mise à jour des paramètres dans le réseau, les rapports quantitatifs seront représentés par la distribution de probabilité conjointe entre les variables modélisées. Cette répartition est décrite de manière efficace en explorant les indépendances probabilistes entre les variables modélisées. Chaque nœud est décrit par une distribution de probabilité conditionnelle sur ses prédécesseurs directs. Nœuds sans prédécesseurs sont décrits par des distributions de probabilité a priori. Par exemple, nœud ***biopsie*** dans le réseau proposé ci-dessous comme le montre la (Fig.4.4) sera décrit par la distribution de probabilité a priori sur ses deux résultats : Bénin et malin

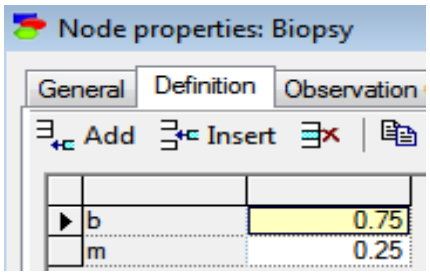


Fig.4. 4 Distribution de probabilité à priori pour le nœud biopsie.

Adénopathie est un nœud qui sera décrite par une distribution de probabilité conditionnelles avec ces prédécesseurs qui sont les nœuds sexe, âge, antécédent familiaux*)*

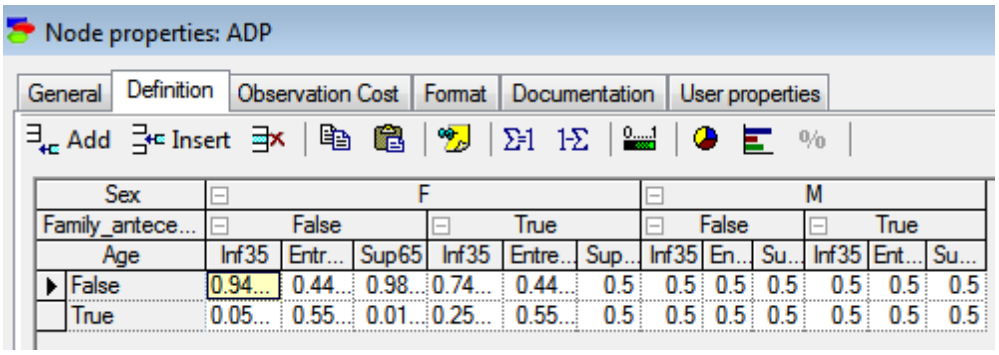


Fig.4. 5 Distribution de probabilité à posteriori pour le nœud Adénopathie.

La structure et les paramètres numériques d'un réseau bayésien peuvent être provoqués par un expert. Ils peuvent également être tirés des données, comme la structure d'un réseau bayésien est simplement une représentation des indépendances dans les données et les chiffres sont une représentation des distributions conjointes de probabilité qui peuvent être déduites à partir des données. Enfin, la structure et les probabilités numériques peuvent être un mélange de connaissances et de mesures d'expert et des données de fréquence objectives.

4.3.4. L'algorithme d'inférence utilisé

Les réseaux bayésiens permettent d'effectuer l'inférence bayésienne, à savoir, le calcul de l'impact de l'observation des valeurs d'un sous-ensemble des variables du modèle sur la distribution de probabilité sur les variables restantes. Par exemple, l'observation d'un ensemble de symptômes, en tant que variables capturées dans un modèle de diagnostic médical, permet pour calculer les probabilités de maladies capturées dans ce modèle.

La mise à jour bayésienne est basée sur les paramètres numériques capturés dans le modèle. La structure du modèle, à savoir une déclaration explicite des indépendances dans le domaine, contribue à rendre les algorithmes de mise à jour bayésienne plus efficace. Tous les algorithmes de mise à jour bayésienne sont basés sur un théorème proposé par le révérend Thomas Bayes (1702-1761) et connu comme le théorème de Bayes.
La mise à jour de réseaux bayésiens est un calcul complexe. Dans le pire des cas, la mise à jour des algorithmes croyances sont NP-difficiles [COO 90]. Il existe plusieurs algorithmes efficaces permettant d'appliquer l'inférence dans les réseaux bayésiens, comme par exemple « Message Passing » [PEA 88], « Junction Tree » [JEN 90], etc. sachant que [WAN 10] a pu montrer que tous les algorithmes d'inférence exacte sur les réseaux bayésiens sont équivalents ou peuvent être dérivés de l'algorithme JLO [DJE 06]. L'algorithme de JLO (clustering) est l'algorithme par défaut de GeNIe et devrait être suffisant pour la plupart des applications. Ce n'est que lorsque les réseaux deviennent très grands et complexes que l'algorithme de clustering peut ne pas être assez rapide. Dans ce cas, il est suggéré que l'utilisateur choisisse un algorithme approximatif.

4.3.4.1. Arbre de jonction (JLO)

Le JLO résout le problème de l'identification du maximum à posteriori (MAP) avec une complexité en temps. La méthode de l'arbre de jonction (clustering ou clique-tree propagation algorithm) a été introduite par Lauritzen & Spiegelhalter et Jensen, Lauritzen & Olesen. Elle

est aussi appelée méthode JLO (pour Jensen, Lauritzen, Olesen). Elle est applicable pour toute structure de GAD contrairement à la méthode des messages locaux. Néanmoins, s'il y a peu de circuits dans le graphe, il peut être préférable d'utiliser une méthode basée sur un ensemble de coupe.

Cette méthode est divisée en cinq étapes qui sont :

- Moralisation du graphe,
- Triangulation du graphe moral,
- Construction de l'arbre de jonction,
- Inférence dans l'arbre de jonction en utilisant l'algorithme des messages locaux,
- Transformation des potentiels de clique en lois conditionnelles mises à jour.

Les étapes sont organisées dans les deux phases suivantes :

- ***La phase de construction*** elle nécessite un ensemble de sous-étapes permettant de transformer le graphe initial en un arbre de jonction, dont les nœuds sont des clusters (regroupement) de nœuds du graphe initial. Cette transformation est nécessaire, d'une part pour éliminer les boucles du graphe, et d'autre part, pour obtenir un graphe plus efficace quant au temps de calcul nécessaire l'inférence, mais qui reste équivalent au niveau de la distribution de probabilité représentée.

Cette transformation se fait en trois étapes :

- La moralisation du graphe,
- La triangulation du graphe et l'extraction des cliques qui formeront les nœuds du futur arbre,
- La création d'un arbre couvrant minimal, appel arbre de jonction ;

- ***la phase de propagation*** **:** il s'agit de la phase de calcul probabiliste à proprement parler où les nouvelles informations concernant une ou plusieurs variables sont propagées l'ensemble du réseau, de manière mettre à jour l'ensemble des distributions de probabilités du réseau. Ceci se fait en passant des messages contenant une information de mise à jour entre les nœuds de l'arbre de jonction précédemment construit. A la fin de cette phase, l'arbre de jonction contiendra à la distribution de probabilité sachant les nouvelles informations, c'est-à-dire P(V/e) où V représente l'ensemble des variables du réseau bayésien et **e** l'ensemble des nouvelles informations sur les dites variables.

1) La phase de construction

a) Moralisation

Définition du graphe moral :

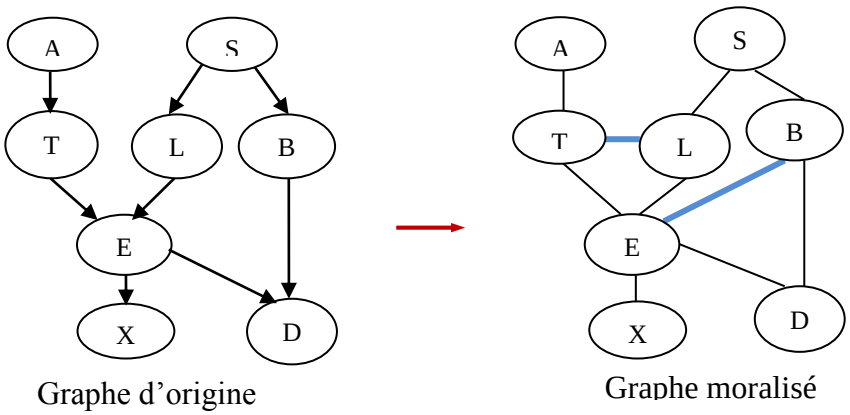
Un graphe moral est un graphe où tous les parents d'un nœud sont connectés par un lien.

La moralisation se décompose suivant les étapes suivantes :

- Mariage des nœuds parents : pour les nœuds possédant plusieurs parents, liaison des parents deux à deux avec des arcs supplémentaires. Récupération du squelette du graphe ainsi obtenu, Nous obtenons alors un graphe non dirigé dit *moralisé*. La (Fig.4.6) montre l'exemple célèbre nommé ASIA par [JEN 96] pour la moralisation d'une partie du graphe proposé dont les huit variables de ce modèle sont :
- A représente une récente vite en Asie,
- T la tuberculose,
- L le cancer du poumon,
- S le fait de fumer,
- B la bronchite,
- D la dyspnée,
- X une simple radio du poumon,
- E variable indicatrice qui prend la valeur 1 lorsque T et L sont positifs

Sachant que A et T sont les variables d'entrées de ce problème et X et D sont des variables de réponses les autres variables permettent l'analyse des données, toutes ces variables sont binaires et aléatoires.

Les arcs colorés représentent les arcs qui ont été rajoutés. La moralisation nécessite que tous les nœuds parents d'un même nœud soient reliés deux à deux.



Graphe d'origine Graphe moralisé

Fig.4. 6 Exemple de la phase de moralisation

b) Triangulation

Définition graphe triangulé :

Un graphe non orienté est triangulé si et seulement si, pour tout cycle de longueur 4 ou plus, il existe une corde, c'est à dire une arête reliant deux nœuds non consécutifs du cycle.

La triangulation consiste à ajouter des liens jusqu'à ce que le graphe soit triangulé. Un graphe peut avoir différentes triangulations qui donnent lieu à des ensembles différents de cliques.

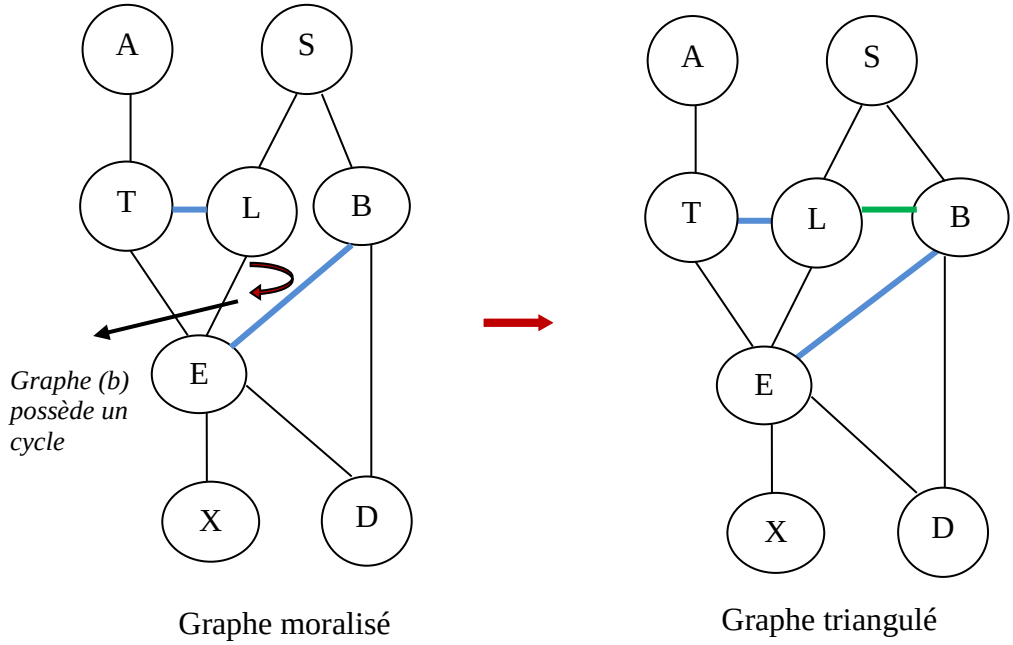
Définition clique :

Une clique d'un graphe non-orienté est un sous-ensemble de sommets de ce graphe dont le sous-graphe induit est complet, c'est à dire que deux sommets quelconques de la clique sont toujours adjacents.

Donc la deuxième étape consiste à trianguler le graphe moral G^m et en extraire des cliques de nœuds, qui sont des sous graphes complets de G. Ces cliques formeront les nœuds de l'arbre de jonction utilisé pour l'inférence. Il faut donc ajouter suffisamment d'arcs au graphe moral G^m afin d'obtenir un graphe triangulé G^T.

L'algorithme de triangulation opère d'une manière très simple. Un graphe est triangulé si est seulement si l'ensemble de ses nœuds peuvent être éliminés. Un nœud peut être éliminé si tous ses voisins sont connectés deux à deux. Donc un nœud peut être éliminé s'il appartient à une clique dans le graphe. Une telle clique forme un nœud pour le futur arbre de jonction qui est en train d'être construit. Ainsi, il est possible de trianguler le graphe et de construire les nœuds de l'arbre de jonction en même temps en éliminant les nœuds dans un certain ordre. Si aucun nœud n'est éliminable, il faut en choisir un parmi les nœuds restants et rajouter les arcs nécessaires entre ses voisins pour qu'il devienne éliminable. Le nœud choisi sera celui pour lequel l'espace d'état de la clique formée sera le plus petit possible. En effet, plus les cliques sont petites, plus l'espace de stockage, et à fortiori le temps de calcul, est réduit.

Donc, suite à notre exemple le graphe obtenu dans la (Fig.4.6) est un graphe non dirigé G^m(moralisé) est dit triangulé si chacun de ses cycles de ***longueur ≥4*** possède un ajout sélectif des arcs au graphe moral pour former un graphe triangulé présenté dans la (Fig.4.7).



Graphe moralisé Graphe triangulé

Fig.4. 7 La phase de triangulation (suite de la figure 6) avec les cliques : C_{TLE}, C_{BLE}, C_{SLB}, C_{DBE}, C_{AT}, C_{XE}

L'efficacité de l'algorithme JLO reste dépendante de la qualité de la triangulation. Mais trouver une bonne triangulation dépend de l'ordre d'élimination des variables. D'une manière générale, trouver une triangulation optimale pour des graphes non-dirigés reste un problème NP-difficile [FOM 13].

c) La construction de l'arbre de jonction

Cette étape consiste à transformer le graphe triangulé en un arbre de jonction. Suite à la triangulation, le graphe est constitué de cliques notées C_i. Notons qu'en général, l'arbre de jonction construit à partir d'un graphe n'est pas unique. Pour cette étape, il suffit de connecter les cliques identifiées, lors de l'étape précédente, avec la condition que toutes les cliques se trouvant dans un chemin entre les cliques C_i et C_j doivent contenir $C_i \cap C_j$. Une fois les cliques adjacentes identifiées, on insère un séparateur, noté S_{ij}, entre chaque paire de cliques C_i et C_j, contenant les variables communes.

Définition de l'intersection courante : Soient C_1 et C_2 deux cliques quelconques de l'arbre de jonction et soit le séparateur $S_{12} = C_1 \cap C_2 = \varnothing$. Alors sur toute chaîne reliant C_1 et C_2, les cliques et séparateurs contiennent S_{12}.

La (Fig.4.8) représente un arbre de jonction associé au graphe de la (Fig.4.6) (réseau

d'origine) contenant quatre cliques (TLE, BLE, AT, SLB, DBE, XE) ainsi que cinq séparateurs (LE,LB,BE,E,T)

i	Les cliques C_i	Séparateur S_i	Les potentiels Φ_i
1	AT	∅	*P (A) P(T|A)*
2	TLE	T	*P (E|T,L)*
3	BLE	L,E	*1*
4	SLB	L, B	*P (S) P (B|S) P (L|S)*
5	DBE	B,E	*P (D|B,E)*
6	XE	E	*P (X|E)*

Tableau.4. 2 des cliques ordonnées

Les cliques du graphe moral triangulé sont données dans le (Tableau.4.2), en les ordonnant suivant un algorithme décrit par [kja 90], cet algorithme serra décrit dans l'exemple d'application suivant. Nous pouvons alors dire que l'ensemble des potentiels de cliques et de séparateurs permettent d'écrire une nouvelle factorisation de la loi jointe sur l'ensemble V des variables du réseau, donnée par :

$$P(V) = \frac{\Pi\ P(Ci)}{\Pi\ P(Si)} = \frac{\Pi\ \Phi(Ci)}{\Pi\ \Phi(Si)} \qquad (4.11)$$

Avec $\Phi_{Ci} = P(C_i)$ et $\Phi_{Si} = P(S_i)$ sont les potentiels de la clique et de séparateur respectivement.

La distribution jointe de cet exemple est :

$$P(V) = \frac{\Phi AT \Phi TLE\ \Phi BLE\ \Phi SLB \Phi DBE \Phi XE}{\Phi T \Phi LE \Phi LB \Phi BE \Phi E}$$

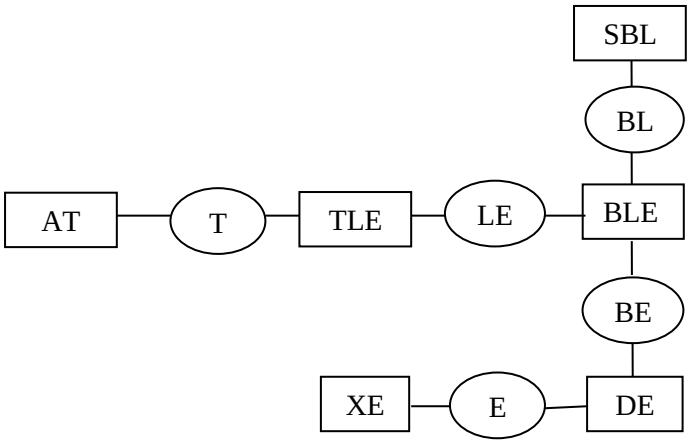


Fig.4. 8 Arbre de jonction associé au graphe G de la fig.4.6

En effet, la dernière clique (XE) aurait pu être voisine de la deuxième clique (TLE) avec comme séparateur E, le choix d'un arbre se fera selon les observations ou selon les probabilités marginales que l'on souhaite calculer. Une fois l'arbre de jonction construit, l'inférence dans le réseau bayésien vont se faire à partir de cet arbre construit grâce à des messages entre cliques.

Ceci permet de réduire considérablement le nombre de variables et donc le nombre d'opérations à effectuer lors de l'inférence.

2) La phase de l'inférence par Propagation des messages

L'inférence probabiliste consiste à calculer la distribution de probabilité *P(V)* ou *P(V|e)* c'est à dire. Avec ou sans observation (évidence e) mais dans l'arbre de jonction on parle de calcule des fonctions potentielsΦ.

Nous résumons le fonctionnement de l'inférence dans les étapes suivantes :

1. **Transformation graphique** qui donne l'arbre de regroupement.
2. **Initialisation** : associer les fonctions des potentiels à l'arbre de regroupement de telle sorte qu'il satisfasse à l'équation (4.11) qui calcule le potentiel des variables V_i de chaque clique en utilisant les probabilités conditionnelles du réseau bayésien $P(V_i|C(Vi))$, les potentiels des cliques et des séparateurs sont initialisés à 1 (l'élément neutre pour la multiplication) :
   - Pour chaque clique et séparateur X on a $\Phi_x(X) \leftarrow 1$
   - Chaque variable V_i de réseau est affectée à une clique avec la contrainte que cette clique doit également contenir les parents de la variable, on multiple

$\Phi_x(X)$ par $P(V_i|C(Vi))$:

$$\Phi_x(X) \leftarrow \Phi_x(X) \times P(V_i|C(Vi))$$

La procédure d'initialisation satisfait à l'équation (4.11) :

$$P(V) = \frac{\prod_{i=1}^{N} \Phi xi}{\prod_{j=1}^{N-1} \Phi Sj} = \frac{\prod_{i=1}^{M} P(Vi|C(Vi)).}{1}$$

Ou V est l'ensemble des variables du réseau, N est le nombre de cliques, M est le nombre de variables et Φ_{xi} , Φs_j sont les potentiels de clique et de séparateur respectivement.

Après l'initialisation de tous les potentiels des cliques par le produit de toutes les probabilités conditionnelles provenant de l'ancienne table de probabilités de réseau bayésien, une nouvelle distribution de probabilités

3. **Propagation** : après avoir initialisé les potentiels de l'arbre de jonction, nous allons exécuter la propagation afin d'avoir une mise à jour consistante de ces potentiels. La propagation exécute une série de manipulations locales appelées **passage de message** dans l'arbre de jonction. On considère deux cliques adjacentes *V* et *W* avec un séparateur *S* et leurs potentiels associés ψ_v, ψ_w et ψ_s. le passage de message de *V* à *W* se produit en deux étapes nommées *la collecte des évidences et la diffusion des messages :*

   - ***La collecte des évidences*** (Mise à jour de *V* à *W* (passage avant) :

     V envoie un message à son parent (racine) *W* en passant par leur séparateur S. Une **projection** du potentiel de la clique *V* vers le séparateur adjacent serra faite en additionnant toute variable qui n'est pas dans le séparateur.

     $$\psi_S^* = \sum_{V/S} \psi_V \qquad (4.12)$$

     Lorsque le parent *W* reçoit un message de son enfant, il multiplie sa table par la table de **message** pour obtenir sa nouvelle table.

     $$\psi_W^* = \frac{\psi_S^*}{\psi_S} \psi_W \qquad (4.13)$$

     Lorsqu'un parent *W* reçoit le message de son enfant, il répète le processus (il agit comme une feuille).

   - ***La diffusion des messages*** (Puis, de *W* à *V* (passage en arrière)) :

     Inverse le passage vers le haut, en commençant à la racine *W.*

     La racine envoie un message à son enfant en passant par leur séparateur une autre **projection de** *W* sur *S* qui se réalise c à d la racine marginalise la table résultante vers le séparateur et envoie le résultat à l'enfant *V.*

$$\psi_S^{**} = \sum_{W/S} \psi_W^* \qquad (4.14)$$

.

L'enfant *V* multiplie sa table par la table de son parent à ce niveau la mise à jour est atteintes.

$$\psi_V^* = \frac{\psi_S^{**}}{\psi_S^*} \psi_V \qquad (4.15)$$

Après un cycle complet de transmission de message, il n'y aura plus de changements dans les potentiels.

Remarque :

On dit que l'arbre de jonction a atteint un état d'équilibre dans la mesure où tous les potentiels des cliques et séparateurs sont cohérents entre eux c'est-à-dire, après les deux interactions de la mise à jour dans l'arbre de jonction tous les potentiels sont consistants [JEN96] :

$$\psi(\mathrm{s}) = \sum_{W/S} \psi(w) = \sum_{V/S} \psi(v)$$

4. **Marginalisation** :

   À partir de l'arbre de jonction complet (tous les potentiels sont actualisés) on calcule P(V) pour chaque variable dans le réseau. A ce moment, obtenir la probabilité *a posteriori* sachant l'évidence d'une variable quelconque, revient à prendre une clique X contenant variable V et à marginaliser la table de probabilités afin d'obtenir la table de probabilité de la variable seule :

   $$P(V) = \sum_{X/V} \Phi \mathrm{x}$$

   Sachant que toute clique contenant la variable d'intérêt est un candidat adéquat [JEN01].

4.3.5. La recherche

À la fin de ce processus, et après que toutes les feuilles (le dernier niveau des variables) il faut mettre à jour leurs probabilités afin d'obtenir leurs probabilités postérieures. Le système (modèle de diagnostic construit sous GeNie) qui a utilisé cet algorithme recherche la variable du dernier niveau avec la probabilité maximale qui sera le résultat du diagnostic des observations (entrées : les signes cliniques, biologique et les attributs de la mammographie):

ArgMaxP (M =maladie | e = évidence).

Exemple d'application :

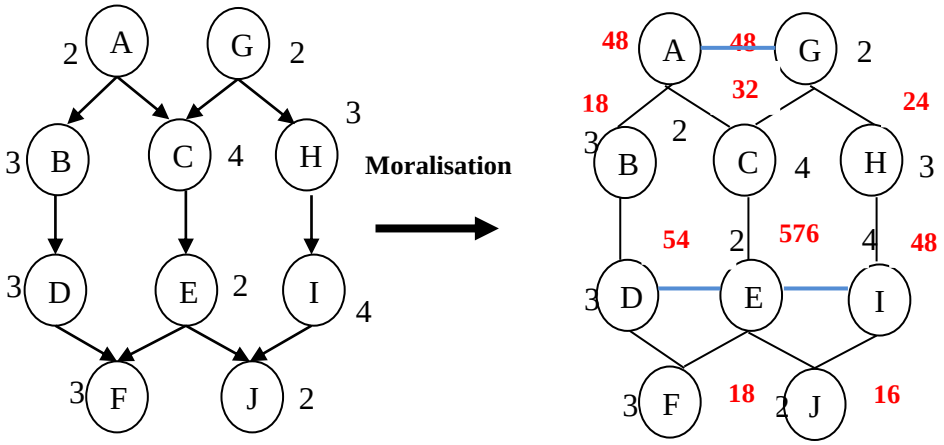
Exemple 1 :

Nous allons voir comment construire l'arbre de jonction ou la triangulation par un exemple, notant que la partie triangulation est la partie la plus difficile car trouver une triangulation optimale (au sens du nombre d'arêtes ajoutées minimum) est NP-difficile. Toutefois en utilisant de bonnes heuristiques, on peut trouver des triangulations qui sont proches en temps polynomial. Comme nous l'avons dit précédemment la qualité de triangulation obtenue dépend uniquement de l'ordre d'élimination α choisi, on peut le résoudre avec un algorithme glouton rapide et efficace « **Kjærulff** » [kja 90] :

Soit un graphe non orienté (moral) G = (X, E), X = {$X_{1, \ldots,}$ Xn}

- Associer à chaque Xi un **poids** égal au produit des modalités (nombre d'état de chaque variable) de X_i et de ses voisins.
- Eliminer le nœud X_i dont le poids est minimal (relier tous ses voisins de manière à former une clique C_i puis éliminer X_i et ses arêtes adjacentes).
- Mettre à jour les poids des nœuds restants.

L'évaluation de l'algorithme de « Kjærulff » donne :

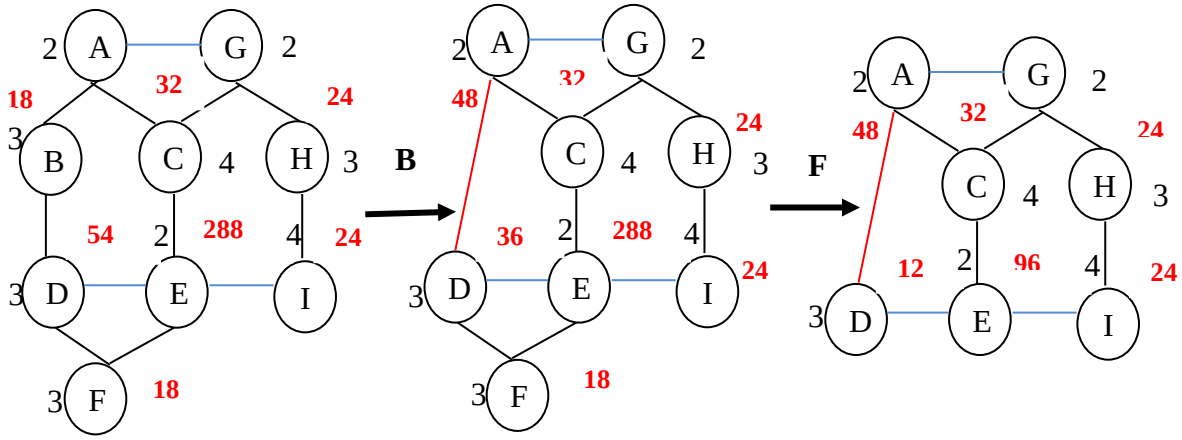


Calculons par exemple le poids de A, nombre de modalités NB_m=2, ces voisins sont : G, B, C leurs NB_m sont respectivement 2, 2, 3, 4.

Donc le poids de A=2*2*3*4=48.

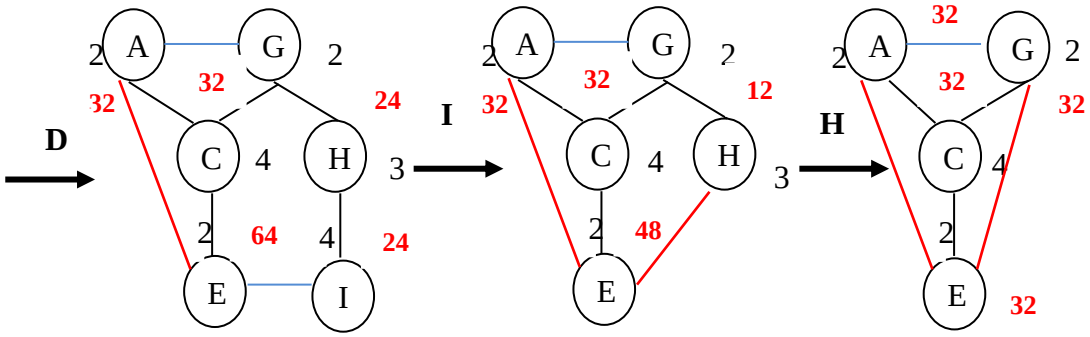
On calcule de la même façon tous les poids des nœuds du graphe moralisé, et à chaque fois le nœud du poids le plus faible serra éliminer.

- Variable à éliminer : J =clique EIJ.



- Première variable à éliminer : B = clique ABD, l'arête à ajouter est : AD
- Deuxième variable à éliminer : F = clique DEF.

On recalcule à chaque élimination les poids.



- Troisième variable à éliminer : D = clique ADE, l'arête à ajouter est : EA
- Quatrième variable à éliminer : I = clique EHI, l'arête à ajouter est : EH
- Cinquième variable à éliminer : H = clique EGH, l'arête à ajouter est : GE
- Sixième variable à éliminer : A = clique ACEG,
- Puis les autres variables peuvent être éliminées dans n'importe quel ordre puisqu'elles appartiennent toutes à la même clique : C = clique CEG

  G = clique EG

  E = clique E

A la fin le graphe non orienté est triangulé, on rajoute des arêtes entre tous les voisins du nœud Xi que l'on veut éliminer (on forme une clique), le graphe triangulé de l'exemple est présenté dans la (Fig.4.9)

0

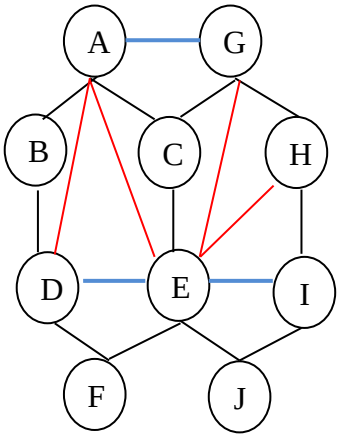


Fig.4. 9 Graphe triangulé par l'algorithme de Kjærulff.

L'ensemble des cliques selon leur ordre de création (avec la variable dont l'élimination a créé la clique) : EIJ (J), ABD (B), DEF (F), ADE (D), EHI (I), EGH (H), ACEG (A), CEG (C), EG (G), E (E).il nous reste à relier les clique selon l'ordre de création, les cliques doivent contenir intersection. Une fois les cliques adjacentes identifiées, on insère un séparateur(S), entre chaque paire de cliques, contenant les variables communes. Ce qui nous permet d'avoir l'arbre de jonction présenté dans la (Fig.4.10) suivant :

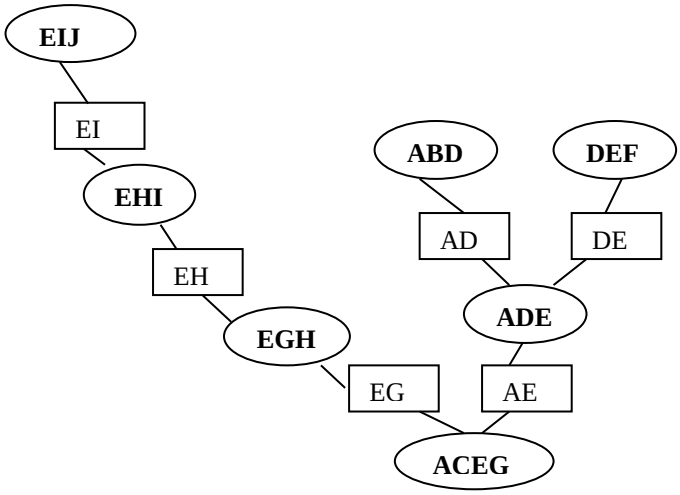
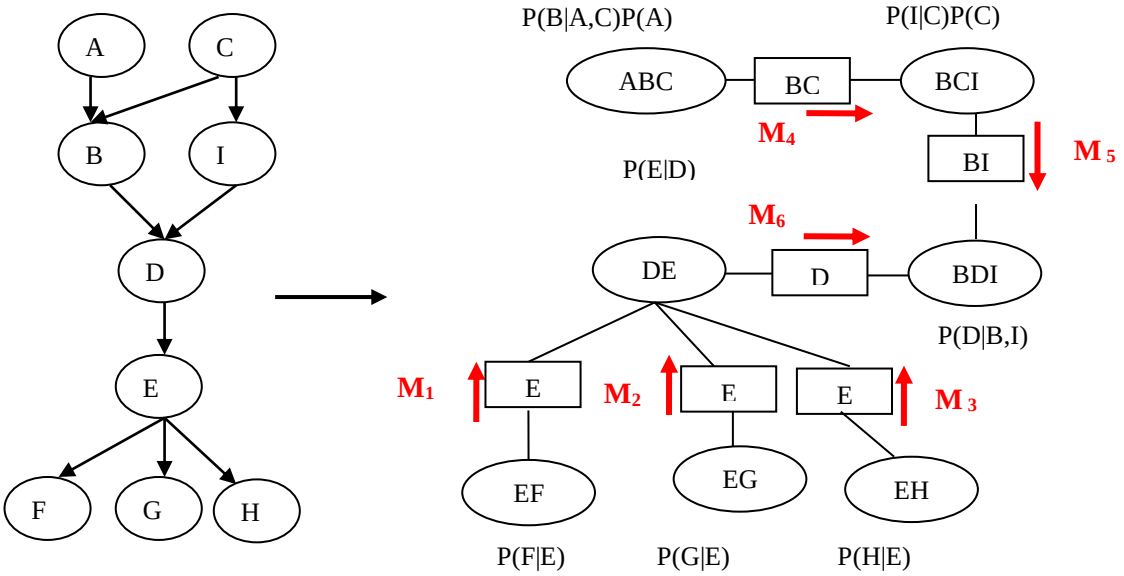


Fig.4. 10 Arbre de jonction.

Exemple 2 :

Nous présentons un autre exemple d'initialisation des potentiels des cliques ainsi les séparateurs, la séquence d'élimination : F, H, G, E, D, I, C, B, A, donne cet arbre de jonction.



Nous présentons la distribution des tables de probabilités associées à ce réseau bayésien

	A_1	A_1
P(A)	0.7	0.3

	C_1	C_2
P(C)	0.6	0.4

		D_1	D_2
P(E/D)	E_1	0.4	0.3
	E_2	0.5	0.4
	E_3	0.1	0.3

		C_1	C_2
P (I/C)	I_1	0.8	0.2
	I_2	0.2	0.8

		A_1		A_2	
		C_1	C_2	C_1	C_2
P(B/A ,C)	B_1	0.5	0.6	0.1	0.8
	B_2	0.5	0.4	0.9	0.2

		B_1		B_2	
		I_1	I_2	I_1	I_2
P(D/B ,I)	D_1	0.1	0.4	0.7	0.8
	D_2	0.9	0.6	0.3	0.2

		G_1	G_2	G_3
P(G/E)	F_1	0.6	0.1	0.3
	F_2	0.4	0.9	0.7

		E_1	E_2	E_3
P (F/E)	F_1	0.2	0.7	0.4
	F_2	0.8	0.3	0.6

		H_1	H_2	H_3
P(H/E)	F_1	0.5	0.6	0.2
	F_2	0.5	0.4	0.8

- **Initialisation** de la table de la clique (ABC) et (BCI)

A	C	B	P (A,B,C)=P(B/A,C)P(A)
a_1	c_1	b_1	0,3*0,5
a_1	c_1	b_2	0,3*0,5
a_1	c_2	b_1	0,3*0,6
a_1	c_2	b_2	0,3*0,4
a_2	c_1	b_1	0,7*0,1
a_2	c_1	b_2	0,7*0,9
a_2	c_2	b_1	0,7*0,8
a_2	c_2	b_2	0,7*0,2

B	C	I	P (B,C,I)=P(I/C)P(C)
b_1	c_1	i_1	0,8*0,6
b_1	c_1	i_2	0,2*0,6
b_1	c_2	i_1	0,2*0,4
b_1	c_2	i_2	0,8*0,4
b_2	c_1	i_1	0,8*0,6
b_2	c_1	i_2	0,2*0,6
b_2	c_2	i_1	0,2*0,4
b_2	c_2	i_2	0,8*0,4

Fig.4. 11 Exemple d'usage de la partie initialisation.

De la même façon pour la clique BCI, BDI, DE, EF, EG, EH. Les séparateurs : BC, BI, D, E seront initialisé à 1

- **Propagation** :

1. *La collecte* : les feuilles envoient des messages M.
   - La clique (EF) se projette sur le séparateur E, on procède de la même façon en calculant les messages M_2 et M_3 pour les cliques (EG) et (EH) respictivement.

$$M_1 := \text{projection}_E\, M(EF) = \text{projection}_E \begin{matrix} & \boldsymbol{e_1} & \boldsymbol{e_2} & \boldsymbol{e_3} \\ \begin{matrix} f_1 \\ f_2 \end{matrix} & \begin{pmatrix} 0{,}2 & 0{,}7 & 0{,}4 \\ 0{,}8 & 0{,}3 & 0{,}6 \end{pmatrix} \end{matrix} = \begin{matrix} \boldsymbol{e_1} & \boldsymbol{e_2} & \boldsymbol{e_3} \\ (1 & 1 & 1). \end{matrix}$$

   - La clique (ABC) se projette sur le séparateur BC, sachant qu'elle est initialisée dans la (Fig. 3.12)

$$M_4 := \text{projection}_{BC} M(ABC) = \text{projection}_{BC} \begin{matrix} & \underline{\quad\boldsymbol{a_1}\quad} & & \underline{\quad\boldsymbol{a_2}\quad} & \\ & \boldsymbol{c_1} & \boldsymbol{c_2} & \boldsymbol{c_1} & \boldsymbol{c_2} \\ \begin{matrix} b_1 \\ b_2 \end{matrix} & \begin{pmatrix} 0{,}15 & 0{,}18 & 0{,}07 & 0{,}56 \\ 0{,}15 & 0{,}12 & 0{,}63 & 0{,}14 \end{pmatrix} \end{matrix} =$$

$$\begin{matrix} \boldsymbol{c_1} & \boldsymbol{c_2} \\ \begin{pmatrix} 0{,}22 & 0{,}74 \\ 0{,}78 & 0{,}26 \end{pmatrix} & \begin{matrix} b_1 \\ b_2 \end{matrix} \end{matrix}$$

   - Le message M_5

$$M_5 := \text{projection}_{BI}\,(M(BCI) \times M4) = \begin{matrix} & \underline{\quad\boldsymbol{b_1}\quad} & & \underline{\quad\boldsymbol{b_2}\quad} & \\ & \boldsymbol{c_1} & \boldsymbol{c_2} & \boldsymbol{c_1} & \boldsymbol{c_2} \\ \begin{matrix} i_1 \\ i_2 \end{matrix} & \begin{pmatrix} 0{,}15 & 0{,}18 & 0{,}07 & 0{,}56 \\ 0{,}15 & 0{,}12 & 0{,}63 & 0{,}14 \end{pmatrix} \end{matrix} \times$$

$$\begin{array}{cc} & \overbrace{\quad\quad\quad}^{b_1} \quad\quad \overbrace{\quad\quad\quad}^{b_2} \\ c_1 \quad c_2 & c_1 \quad\quad c_2 \quad\quad c_1 \quad\quad c_2 \end{array}$$

$$\begin{pmatrix} 0{,}22 & 0{,}74 \\ 0{,}78 & 0{,}26 \end{pmatrix} \begin{matrix} b_1 \\ b_2 \end{matrix} = \begin{pmatrix} 0{,}1056 & 0{,}0592 & 0{,}3744 & 0{,}0208 \\ 0{,}0264 & 0{,}2368 & 0{,}0963 & 0{,}0832 \end{pmatrix} \begin{matrix} i_1 \\ i_2 \end{matrix} =$$

$$\begin{array}{cc} b_1 & b_2 \end{array}$$

$$\begin{pmatrix} 0{,}1648 & 0{,}3952 \\ 0{,}2632 & 0{,}1746 \end{pmatrix} \begin{matrix} i_1 \\ i_2 \end{matrix}$$

M_6 := projection$_D$(M(ED)×M_1×M_2×M_3)= projection$_D$(P(E|D)×M_1×M_2×M_3).

2. *La diffusion* : la diffusion commence lorsque la clique racine (BDI) reçoive tous les messages des feuilles, en ce moment-là, la diffusion commence à partir de la clique (BDI) par une transmission des messages aux cliques adjacentes en se projetant par ses séparateurs. La procédure se poursuit de la même manière de sorte que chaque passage de message préserve la mise à jour présentée par les passages de messages précédents. Une fois la propagation est terminée, chaque paire de clique-séparateur est mise à jour et l'arbre de jonction et ainsi complet.

- **Marginalisation**

Après avoir actualisé tous les potentiels (toutes les feuilles reçoivent les messages transmis de la racine) de l'arbre de jonction présenté l'exemple 2, nous présentons un exemple de marginalisation de la clique (ABC) qui présente une feuille pour la partie diffusion montrée dans la (Fig.4.12)

A	C	B	Φ_{ABC}(ABC)
a_1	c_1	b_1	0,090
a_1	c_1	b_2	0,090
a_1	c_2	b_1	0,072
a_1	c_2	b_2	0,048
a_2	c_1	b_1	0,042
a_2	c_1	b_2	0,378
a_2	c_2	b_1	0,224
a_2	c_2	b_2	0,056

P(B)= $\sum_{AC} \Phi_{ABC}$:

B	P(B)
b_1	0,090+0,072+0,042+0,224=0,428
b_2	0,090+0,048+0,378+0,056=0,572

Fig.4. 12 Exemple de marginalisation

4.4. Maintenance du système de diagnostic médical

La maintenance met en œuvre des politiques de développement pour la mise à jour aux représentations de la structure à un système de diagnostic et se concentre sur son organisation pour faciliter le futur raisonnement rencontré [REF 11]. Cette maintenance développe des techniques de contrôle et de réponse aux changements de ces différentes sources de connaissances. À mesure que le système fonctionne, il peut provoquer des erreurs ou il peut produire des solutions inadéquates. S'il ne peut pas apprendre de ces erreurs, le système perd de son intérêt et de son importance. Il s'avère donc, nécessaire de maintenir et de mettre à jour la structure et les données pour permettre l'apprentissage du système.

4.5. Maintenance pour quels types d'applications ?

Pour ce type d'études, l'application la plus courante est l'aide au diagnostic, très utilisée en médecine. Ainsi, si les variables du modèle sont suffisamment nombreuses et décrivent bien le processus, il est alors possible de déterminer le défaut le plus probable du système.
Partant de là, l'outil GeNie peut également faire l'aide au diagnostic. En effet, les experts peuvent donner des variables modélisant les actions de maintenance. Ces actions vont être intégrées dans le réseau bayésien en tant que variables toujours observées. De nouveau, les experts devront être capables de quantifier le bénéfice de la maintenance.

4.6. Maintenance de réseau bayésien

Définition :

Maintenir un réseau bayésien revient à définir des variables supplémentaires notées comme des actions de maintenances pour une mise à jour du modèle graphique ou on pourrait les ajoutés ou supprimés et modifiés leurs valeurs. Ces variables sont toutes des variables d'entrée. En effet nous considérons que ces variables susceptibles d'agir sur les variables du modèle, mais que l'inverse n'est pas possible. De plus, comme c'est le cas dans notre étude, les actions de maintenance influent sur les paramètres de sortie, mais également sur une variable spécifique choisit par les experts du domaine qui est le cancer des seins.
L'établissement d'un réseau bayésien se heurte à la multiplicité des probabilités à priori et surtout les probabilités conditionnelles à renseigner par les experts il suffit de quelques variables comportant de nombreux liens pour rendre cette tache indispensable, et réalisable [COR 03].

4.6.1. Maintenance de la structure réseau bayésien

Nous proposons la maintenance du réseau bayésien correspond au modèle illustré par la (Fig.4.2). Les actions de maintenance permettent de maintenir la base de données et peuvent être considérées comme des variables du modèle graphique.

En outre, le réseau bayésien est composé d'un ensemble de variables liées par des relations de dépendances entre elles. Ces variables se situent en amont des variables sur lesquelles elles agissent. Enfin, des probabilités conditionnelles doivent être également données par les experts.

Objectif :

Dans notre étude, l'objectif est de retenir des variables critiques efficaces afin de fournir des tâches de maintenance sur ces variables du modèle. Ces variables ont été choisies par les experts. Ensuite, les experts ont sélectionné des tâches de maintenance sur les variables et donc sur les composants choisis. Ces actions de maintenances peuvent être vues comme des variables du modèle graphique [REF 12].

Enfin, la probabilité conditionnelle doit être donnée par l'expert. L'un des objectifs consiste à repérer les politiques de maintenance pour augmenter la probabilité et prendre en compte le temps de réponse et la représentation de la base de données [COR 03].

La maintenance du réseau bayésien (Fig.4.13) prend en compte la collecte des actions données par l'expert du domaine, ainsi que le réseau initial. Cette phase vise à décider si nous intégrerons ou non les différentes connaissances véhiculées par les règles de maintenance.

Ainsi, la définition d'une nouvelle action dans notre modèle nécessitera l'opinion de cet expert. L'ensemble des actions ont été pensés pour faciliter leurs interprétations en vue d'une intégration au modèle (ajout, modification ou suppression) d'arcs ou de nœuds du graphe. Le cas échéant, il faut redéfinir les tables de probabilités impactées par les modifications, il est possible de s'appuyer sur les données et les règles de l'expert du domaine.

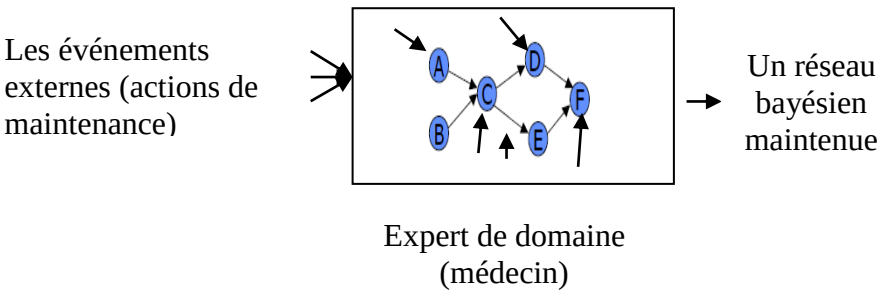


Fig.4. 13 Maintenance de la structure

4.6.2. Choix des actions de maintenance

L'interface de travail GeNIe comporte divers éléments comme la barre d'outils standards qui permet d'ajouter ou de supprimer des nœuds et des arcs, ce dernier facilite la tâche de maintenance pour les experts. À ce niveau, les actions de maintenance sont de type (entrée, intermédiaire).

Pour notre problème (des maladies des seins) nous présentons les facteurs de risque comme étant des actions de maintenance concernant la maladie « cancer du sein » [REF 11] graphiquement présentée dans le réseau bayésien ainsi leurs probabilités a priori sont conditionnelles, sachant qu'elles sont binaires et leurs types sont discrets, et qui sont comme suit : (A: désigner une action de maintenance).

- A_1 : Obésité, avec probabilité P (A_1).
- A_2 : Prise de pilules, avec probabilité P(A_2).
- A_3 : L'âge d'avoir des enfants P(A_3).
- A_4 : La consommation de tabac P(A_4).
- A_5 : Exposition à des rayonnements ionisants
- A_6 : Une ménopause tardive.
- A_7 : Une trop faible consommation de fruits et légumes.

Une action de maintenance peut concerner une ou plusieurs variables. Mais plusieurs actions de maintenance ne peuvent concerner la même variable. Autrement dit, nous traitons le cas de la présence d'une seule action de maintenance ou le cas de deux actions de maintenance agissant sur la variable jugée importante « cancer du sein ».

4.6.3. Intégration des actions de maintenance

Le choix d'intégrer les actions de maintenance est une phase très importante. Il rend le système de diagnostic plus flexible et donne une conception dynamique de notre graphique. Le schéma (Fig.4.14) ci-dessous décrit les étapes de diagnostic développées dans les paragraphes précédents avec la phase de maintenance qui est facultative et qui concerne l'expert du domaine.

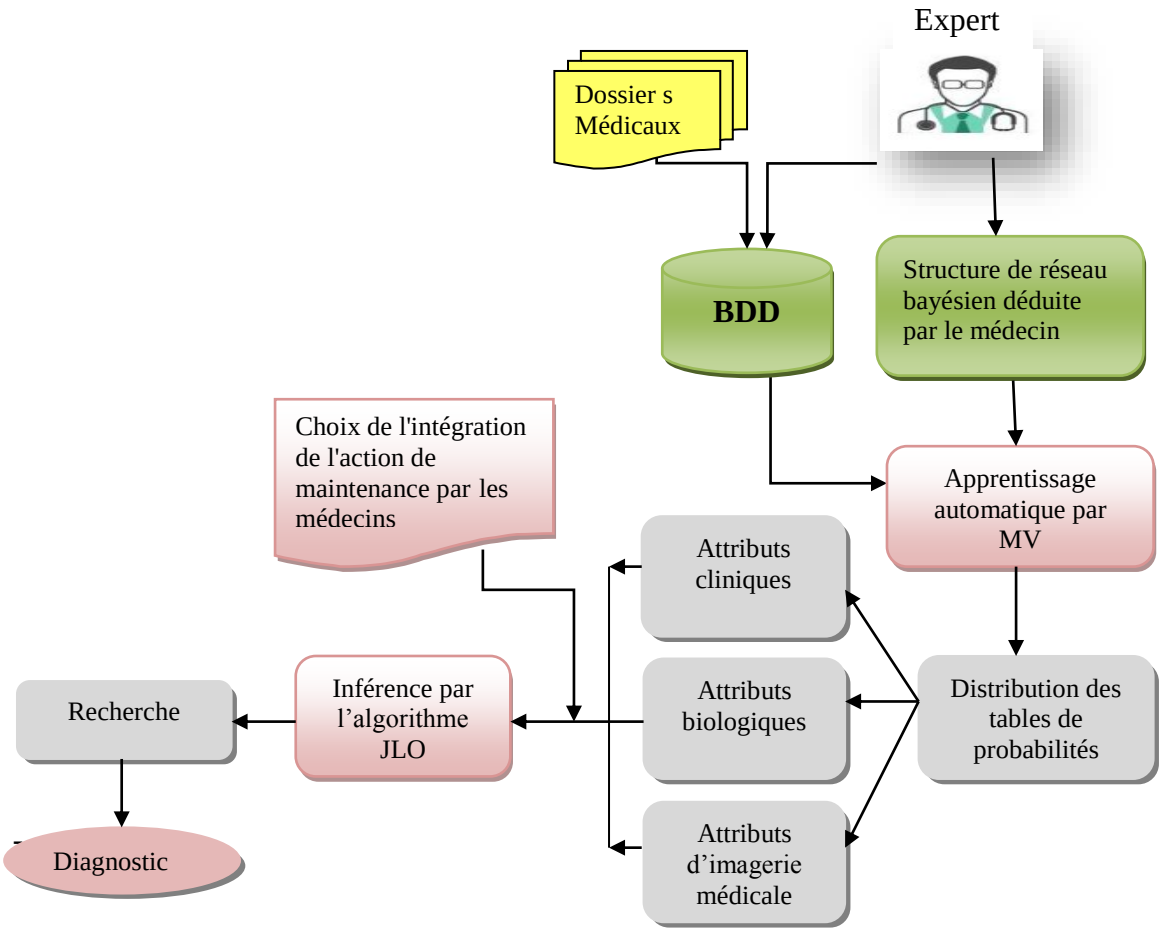


Fig.4. 14 Description de l'architecture du système de diagnostic

Le réseau bayésien est présenté à la (Fig.4.15) avec les actions de maintenance. Une action de maintenance peut également influencer une ou plusieurs variables et non nécessairement des variables de même type (entrée, intermédiaire, sortie). Une variable peut également être influencée par une ou deux actions de maintenance. Cette dernière situation est rencontrée dans notre étude, Donc nous avons présenté, pour illustrer ce que nous avons étudié, un exemple avec deux paramètres d'ajout sur notre réseau bayésien résumé dans (Fig.4.15) et un exemple d'application, que nous allons voir dans le chapitre suivant leur influence sur la maladie concerné (le cancer des seins).

Une fois cette variable importante identifiée, l'aide à la décision prend part via l'intégration des actions de maintenance sur cette variable jugée importante. Pour cela nous considérons les taches de maintenance comme de nouveau nœuds de réseau bayésien apportés avec leur probabilité conditionnelle, une nouvelle inférence est faite avec la prise en compte des actions de maintenance. Les résultats apportent une augmentation de probabilité d'apparition de cancer des seins par rapport au résultat.

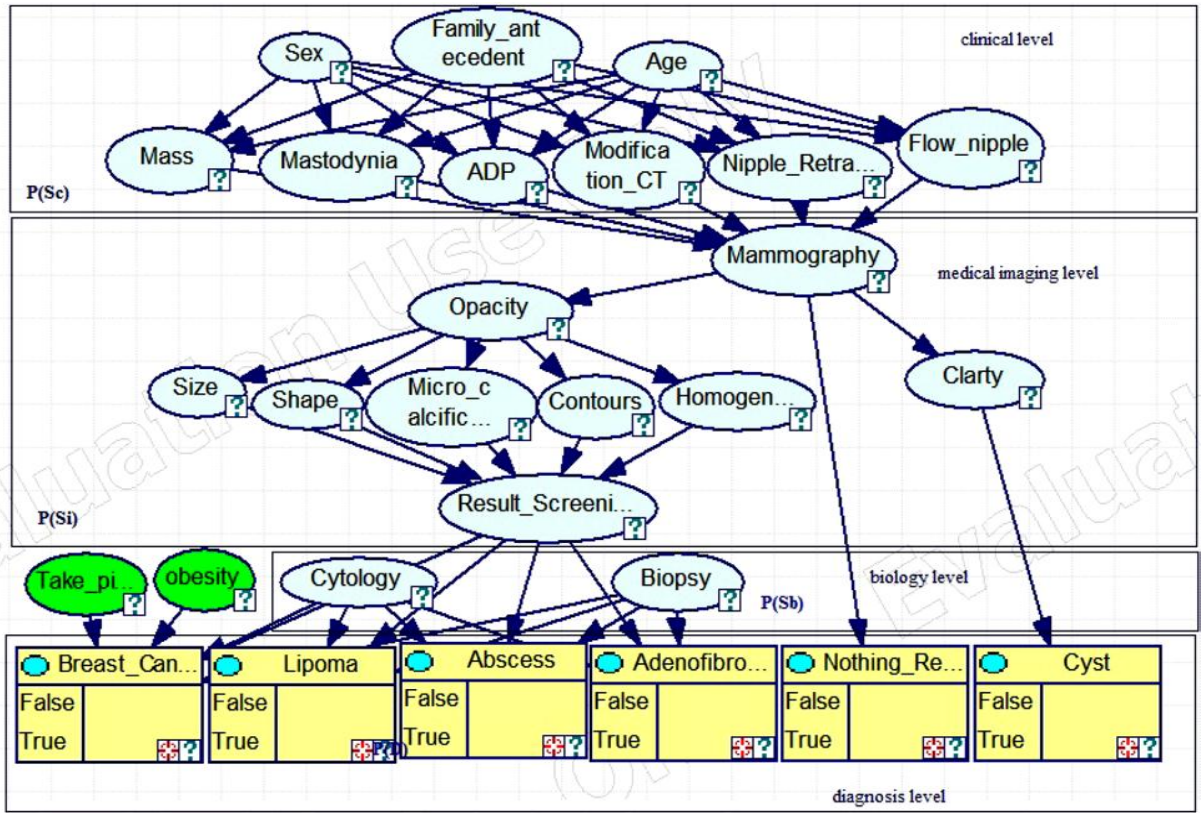


Fig.4. 15 Diagramme général du système de modélisation avec deux actions de maintenance.

4.7. Conclusion

Au cours de ce chapitre nous avons fait une analyse du problème que nous avons spécifié dans le processus de diagnostic en mettant l'accent sur le formalisme des réseaux bayésiens. Puis nous avons présenté l'exploitation du modèle proposé pour l'aide au diagnostic des maladies les plus fréquentes des seins par l'outil bayésien GeNIe.

Nous avons détaillé la conception de chaque étape pour la réalisation d'un système d'aide au diagnostic par l'implémentation des deux partie qualitatif et quantitatif du réseau bayésien sous GeNIe ainsi, nous avons présenté l'algorithme choisi pour la tâche d'inférence qui combine ces deux parties par une démarche logique probabiliste.

En effet, chaque système informatique notamment les systèmes qui essayaient de simuler les raisonnements des médecins incertains nécessite en phase de fonctionnement d'être maintenu pour garantir la qualité, la performance, et la compétence de ce système. Donc, nous avons présenté la maintenance du réseau bayésien qui donne la possibilité d'intégré des actions de maintenance comme variable du modèle afin de remédier le problème de la mise en place d'un seul modèle de réseau bayésien.

Chapitre 5 Implémentation du système SADMseins

5.1. Introduction

La construction des modèles de réseau bayésien est une tâche complexe et longue. Il est difficile d'obtenir des modèles complets et cohérents, mais obtenir des données de probabilités correctes et fiables pour les modèles conçus est beaucoup plus difficile.

Ce chapitre vise à décrire les fondamentaux et les techniques de mise en œuvre d'un modèle de réseau bayésien dont Il existe de nombreuses implémentations dans une variété de formats et de langues. Nous avons opté pour cela la combinaison du logiciel GeNIe [DRU 99] avec la bibliothèque SMILE qui fournit un moyen simple de développement polyvalent et convivial pour les modèles de décision graphiques et de diagnostic.

Cette partie cernera le matériel employé, et aussi les différents modules qui composent l'outil bayésien choisi, nous abordons toutes les étapes de la validation du modèle graphique qu'elles nous ont amenées à l'évaluation du système **SADMseins** et sa maintenance.

5.2. Spécification de l'outil de développement (SMILE/GeNIe)

Il existe des outils qui implémentent les algorithmes de propagation des résultats de nouvelles preuves via les réseaux bayésiens, ainsi que la fourniture d'une interface graphique à l'utilisateur pour dessiner les graphes et compléter les tables de probabilités.

SMILE (Structural Modeling, Inference, and Learning Engine) est une bibliothèque entièrement portable de classes C++. Il permet l'implémentation des méthodes basées sur la théorie de la décision comme les réseaux bayésiens et les diagrammes d'influence. Donc SMILE est un moteur de raisonnement utilisé pour les modèles probabilistes graphiques et fournit des fonctionnalités pour effectuer un diagnostic. GeNIe

L'aspect important de SMILE est l'existence des adaptateurs de la bibliothèque pour d'autre langage de programmation comme : jSMILE pour Java, SMILEX pour ActiveX et SMILE.NET.

Le nom du GeNIe et sa capitalisation inhabituelle proviennent du nom de l'interface de réseau graphique,

- Il constitue l'interface graphique de la bibliothèque SMILE.

- Il est implémenté dans visual C++ et s'appuie fortement sur MFC (Microsoft Foundation Classes).
- Il permet de construire des modèles de n'importe quelle taille et complexité.
- Limitée par la capacité de la mémoire d'exploitation de l'ordinateur.
- Les modèles développés en utilisant GeNIe peuvent être intégrés dans toutes les applications et exécutés sur toute plate-forme informatique, en utilisant SMILE, qui est entièrement portable.
- L'utilisateur devrait avoir des connaissances approfondies sur les modèles graphiques probabilistes.
- Il dispose d'un guide : aide en ligne.

5.3. Les exigences matérielles et logicielles

Espace disque

L'installation complète de GeNIe nécessite moins de 20 Mo d'espace disque.

Mémoire

Les tables de probabilités conditionnelles croîts de façon exponentielle avec le nombre de parents d'un nœud. Le nombre maximal de parents d'un nœud déterminera les besoins en mémoire. En outre, les besoins en mémoire de l'algorithme de clustering se développent avec la connectivité du réseau

Système opérateur

GeNIe fonctionne sous les systèmes d'exploitation de la famille Windows (98 / NT / 2000 / XP, vista, windows7 …). Les utilisateurs de GeNIe à l'échelle mondiale ont également déclaré avoir exécuté GeNIe sur Linux et Mac OS en utilisant Wine.

Version GeNIe

La version que nous avons utilisé dans ce travail est GeNIe 2.1 représenté dans la (Fig.5.1). Pour déterminer la version de GeNIe que nous avons installée, sélectionnez **À propos de GeNIe** dans le menu **Aide**. Le numéro de la version est répertorié dans le petit cadre de la fenêtre suivante :

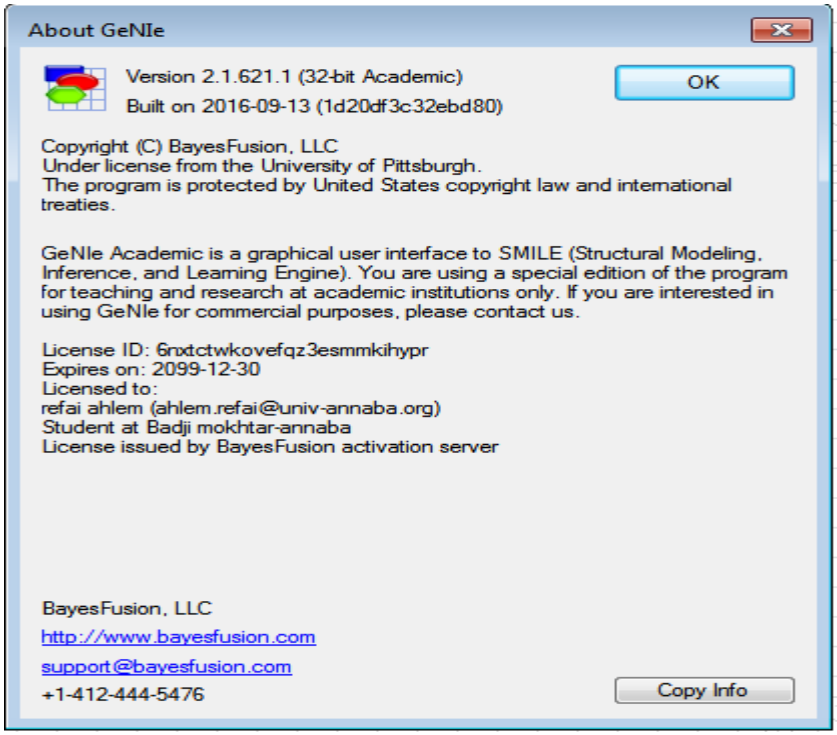


Fig.5. 1 fenêtre déterminant la version de GeNIe utilisé

5.4. Interface principale de GeNIe

L'interface principale est l'environnement du travail de GeNIe. Son objectif principal est de nous permettre de visualiser le réseau en cours de développement de plusieurs façons alternatives ; Les vues fondamentales auxquelles la plupart des utilisateurs travaillent sont la vue graphique et la vue arborescente. Cette interface se divise en quatre parties présentées dans la (Fig.5.2) :

A. La zone des commandes (la barre de menu, format toolbar, la barre d'outils et l'icône de contrôle) qui intègrent l'ensemble des commandes pouvant intervenir sur le graphe actif.

B. La zone de la vue graphique dans laquelle s'ouvrent des fenêtres de graphes.

C. La zone de la vue arborescente, affiche les relations entre les éléments du réseau à l'aide d'une structure arborescente.

D. La barre d'état présente une courte description de la commande à exécuter par l'élément de menu sélectionné, elle répertorie le nombre de nœuds de preuves et les nœuds cibles présents dans le réseau actif.

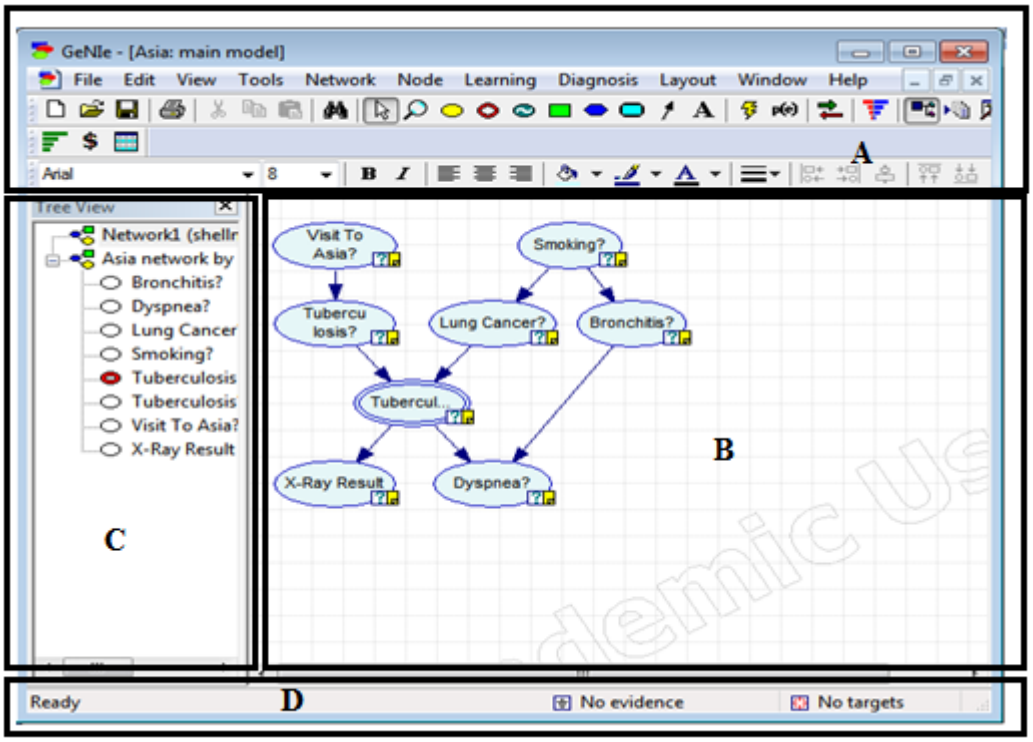


Fig.5. 2 L'interface principale de GeNie

5.5. Étude de cas

Nous proposons une approche probabiliste qui utilise les réseaux bayésiens pour l'aide au diagnostic des pathologies des siens. Les principales étapes sur lesquelles repose l'approche sont les suivantes :

1- Construction de la structure du modèle proposée à l'aide de l'expert (médecin) en l'implémentant graphiquement sous GeNIe.
2- Collecter des dossiers médicaux par l'étude statistique à l'hôpital Ibn Roched d'ANNABA.
3- Affectation des paramètres (tables de probabilités des nœuds).
4- Inférence bayésienne (algorithmes, outils).
5- Résultats (la recherche de diagnostic avec et sans la maintenance bayésienne).

5.5.1. Création du réseau bayésien

Un réseau bayésien peut être construit soit à partir d'un ensemble de données, soit à partir de jugements d'experts, soit à partir d'une combinaison des deux. Nous avons opté pour cela une combinaison dont la structure du réseau est décrite avec l'aide des médecins et les probabilités proviennent de l'apprentissage automatique à partir d'une base de données.

La structure d'un réseau bayésien est une illustration graphique, *les données sont de type qualitatif* : la création de la structure du graphe (Nœud et Arc) se fait dans la zone quadrillée de la vue graphique.

Nœud :

Le menu **Outil** montre une liste des différents types de nœuds que nous pouvons créer. Pour notre cas nous avons utilisé le type de nœud ***chance*** qui comporte 3 types de bases : Général, Noisy Max et Noisy Adder. La spécification des nœuds du réseau proposé dans notre travail est de type : ***chance-général***.

Les nœuds de chance, dessinés comme ovales, désignent des variables aléatoires sachant qu'on a trois types de nœuds de diagnostic qui peut prendre comme état : l'état cible (target ()), auxiliaire, Observé (évidence), comme ils peuvent prendre aussi avant et lors de l'exécution l'état Contrôlé (), Invalide (), valide et Implicite ().

Si on veut dessiner un nœud de type chance on doit procéder selon les étapes suivantes :

1. Sélectionnez le bouton Chance à partir de la barre d'outils standard ou à partir du menu outil (tools).
2. Déplacez la souris pour une partie claire de la zone graphique.
3. Faire glisser le curseur, un nouveau nœud sera créé.
4. Définition les propriétés du nouveau nœud : identifiant, le nom.

On répète les mêmes étapes pour créer plus de nœuds. Tous les nœuds doivent être des nœuds de type ***Chance*** pour le diagnostic (Fig5.3).

Arc :

Pour ajouter un arc entre deux nœuds,

1- Sélectionner l'outil de l'arc de la barre d'outils standard et cliquer avec le bouton gauche sur le nœud parent.
2- Faites glisser le curseur de la souris sur le nœud enfant puis on relâche le bouton de la souris.

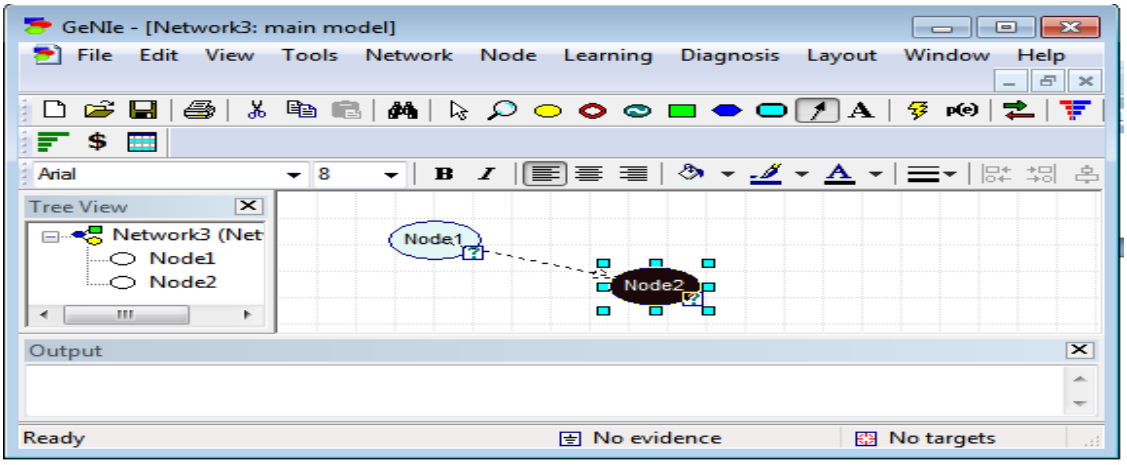


Fig.5. 3 Création du réseau bayésien

La suppression d'un nœud/arc

Dans ce mode, il suffit de sélection l'objet (nœud ou arc) puis en cliquant sur le bouton droit de la souris et on choisit la commande ***delete*** ou directement du bouton ***suppr*** du clavier l'objet sera supprimer.

La modification de la variable

Dans ce mode, on peut modifier toutess les propriétés d'un nœud tel que : identifiant, nom, les états, les valeurs en cliquant deux fois sur le nœud qui fait apparaitre la fenêtre : ***node properties*** (Fig.5.4)**.**

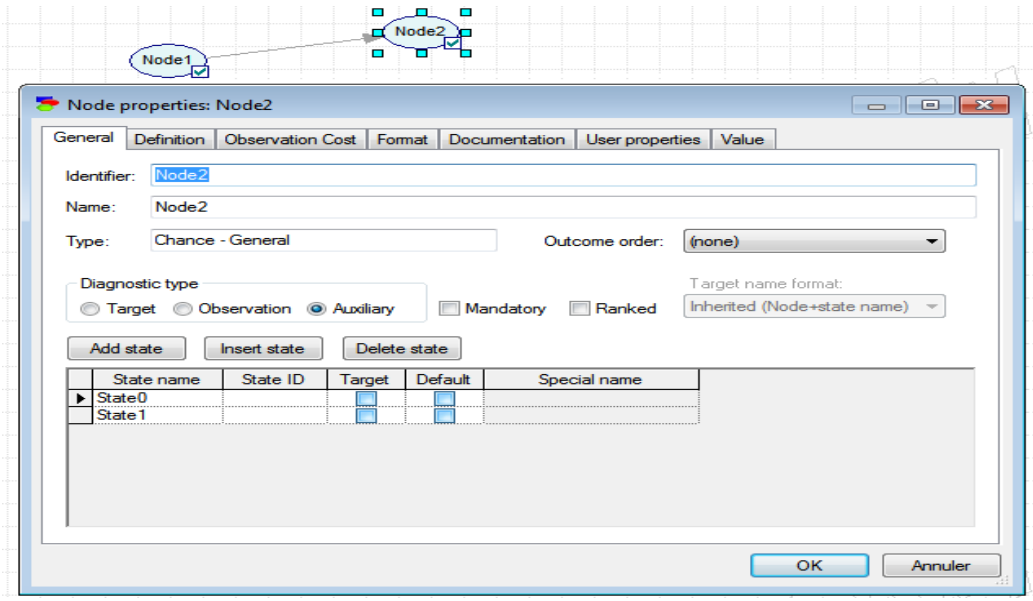


Fig.5. 4 fenêtre propriété du nœud

5.5.2. La collection des données :

Nous avons obtenu la BBD en s'appuyant sur les principales sources d'acquisition suivantes :

- La Collecte des dossiers médicaux par l'étude statistique à l'hôpital Ibn Roched d'ANNABA avec l'assistance du Professeur Aouras Hayet.
- La consultation des sites web médicaux tels que le site : http://www.cancersein.net.
- Des dialogues avec les médecins et les radiologues.

5.5.3. Affectation des paramètres

Les paramètres représentent les données quantitatives : Table de probabilités des nœuds de réseau bayésien. Nous avons estimé les distributions de probabilités à partir d'une base de données comporte 100 enregistrements.

GeNIe peut accéder à des données provenant de trois sources : les fichiers texte, les bases de données ODBC et le format de données GeNIe natif. La base de données ODBC fera l'objet de la section suivante.

ODBC (Open Database Connectivity) est une interface de programmation d'application standard pour accéder aux systèmes de gestion de bases de données (SGBD). GeNIe implémente la norme ODBC, ce qui lui permet de se connecter à la plupart des SGBD.

Pour apprendre les paramètres d'un réseau bayésien existant (c'est-à-dire celui pour laquelle la structure est déjà défini), nous avons besoin d'un fichier de données et d'un réseau ouvert.

Dans cette section, nous ouvrirons une base de données Microsoft Access. Pour accéder aux données d'une base de données, on sélectionne le menu **File →import ODBC data**, qui ouvrira la fenêtre **« sélectionner la source de données » →source de données machine**, GeNie affichera une boîte de dialogue qui permet de sélectionner le fichier de données Access **« bd_maladie_sein** », qui devrait ressembler à (Fig.5.5) :

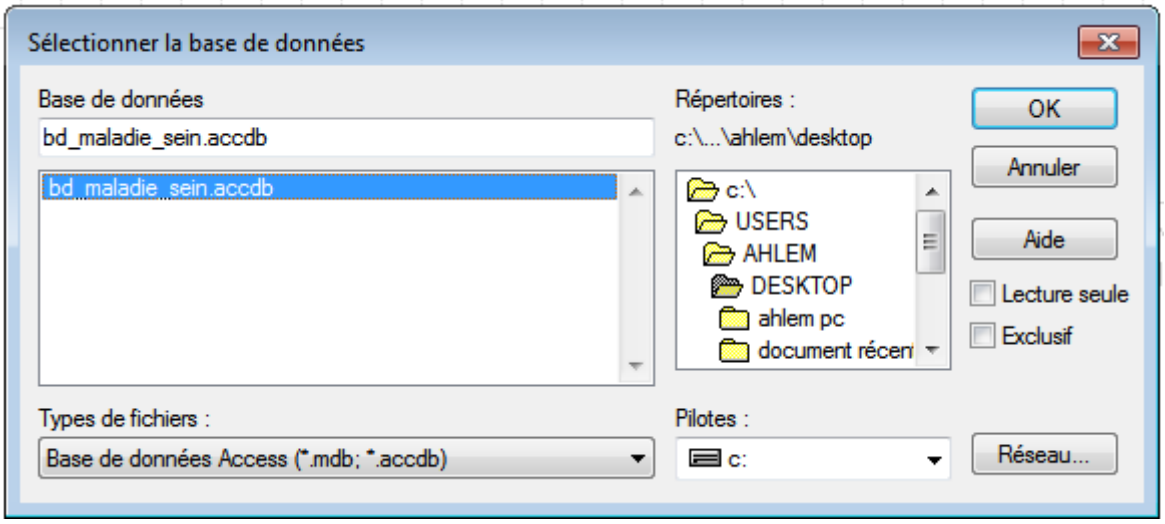


Fig.5. 5 l'import de la base de données

On clique sur le fichier de données **« feuil1 »** importé puis **ok** GeNIe utilise la vue grille pour afficher le contenu d'un fichier de données présenté dans la (Fig.5.8). La grille est semblable à une feuille de calcul, comme Microsoft Excel, et permet aux utilisateurs d'analyser et de nettoyer les données. Le bas de la fenêtre affiche le nombre de lignes dans lesquelles réside le curseur, ainsi que le nombre total de lignes.

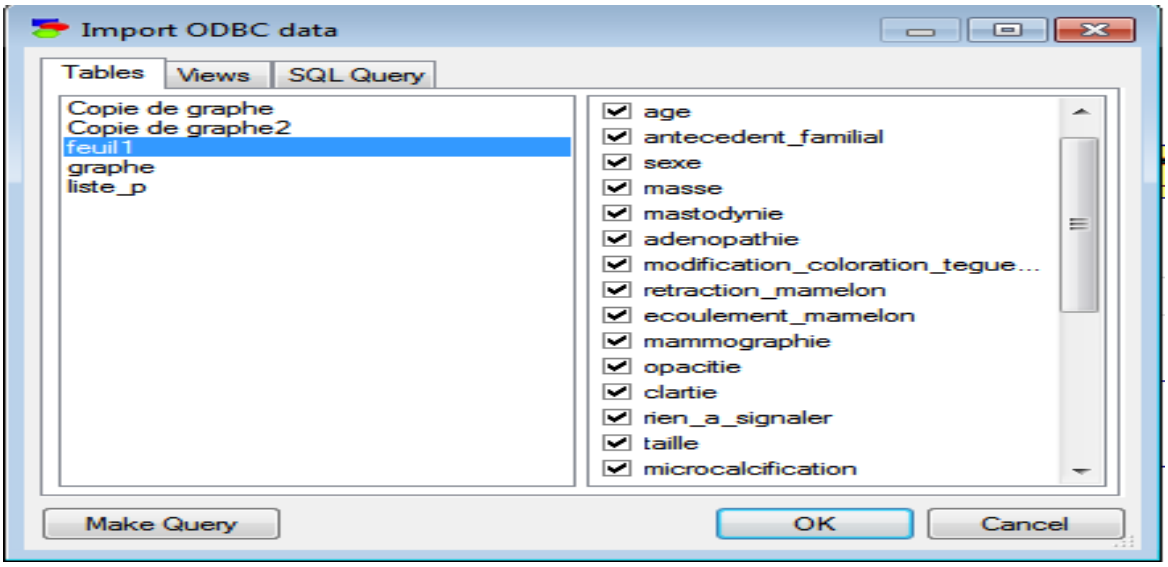


Fig.5. 6 présentation de toutes les variables importées par GeNIe

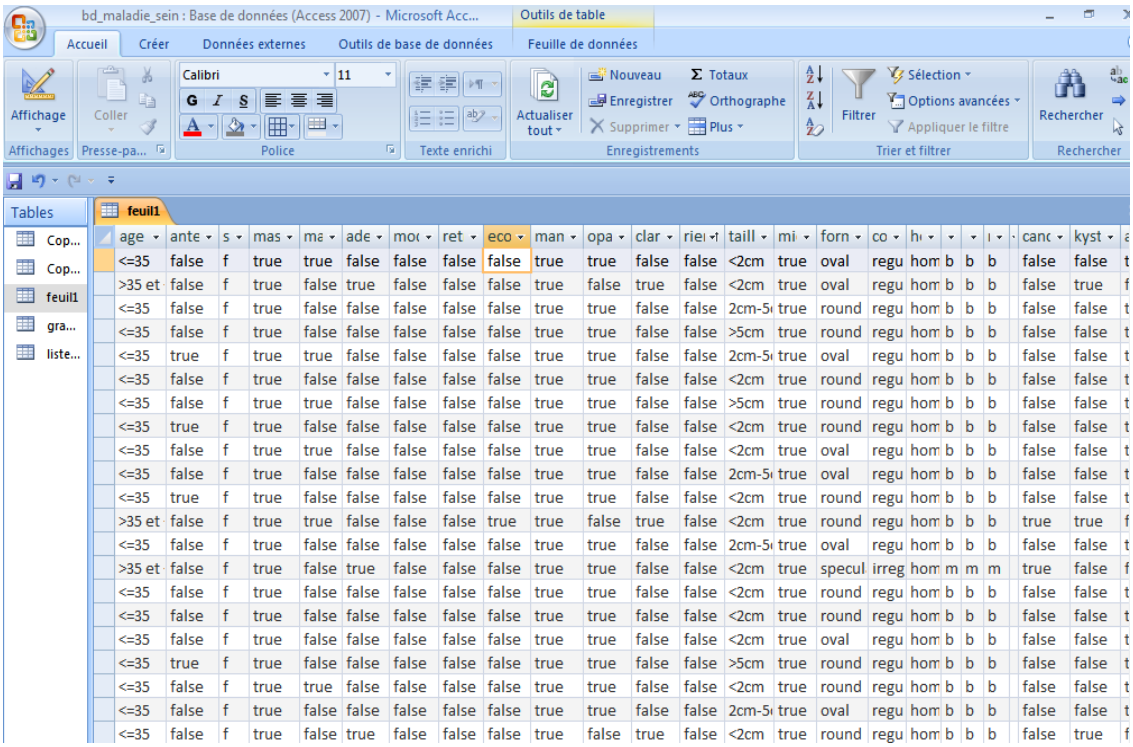


Fig.5. 8 La fiche de la base de données présentée sous Microsoft Access

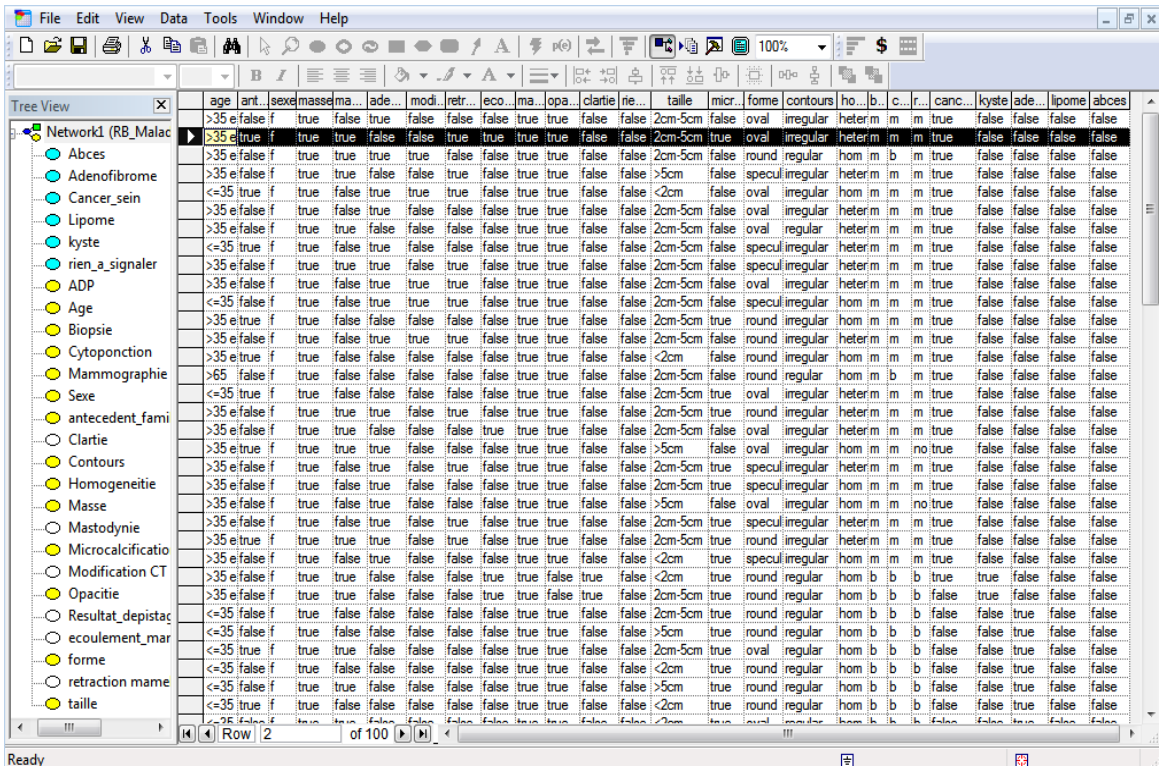
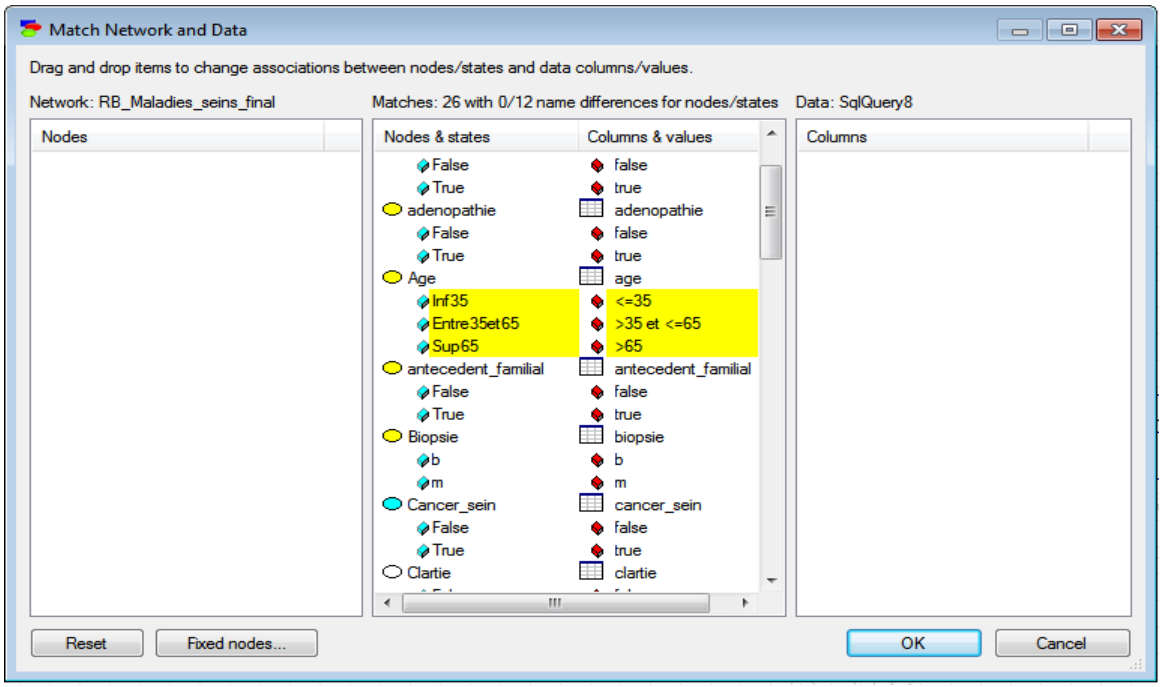


Fig.5. 7 La fiche de la base de données présentée sous GeNie

Les colonnes contiennent des variables (leurs identifiants sont dans la première rangée grisée), les lignes contiennent des enregistrements de données, une fois que nous avons ouvrir la base de données nous évaluons **le modèle** avec **le fichier de données** en cliquant sur le menu **learning → learning parameters** permettra de réafficher la fenêtre (Match Networkand Data), dont la seule fonction est de s'assurer que les variables du modèle (colonne gauche) sont désigné précisément aux variables définies dans la base de données (colonne droite).La (Fig.5.9) présente une disparité dans la variable **Age du réseau** et la variable **Age de la fiche de données**.



-

Fig.5. 9 Fenêtre de la correspondance entre le réseau et les données

Sachant que les variables définies dans le réseau et les variables définies dans l'ensemble de données doivent avoir des noms identiques ou proches, si ce n'est pas le cas une fenêtre apparaisse **« match network and data »** comporte les deux listes de variables qui sont classées par ordre alphabétique. S'il existe une disparité entre eux, GeNIe souligne les différences avec un fond jaune, ce qui facilite l'identification des disparités. La correspondance manuelle entre les variables du modèle et les données sera effectuée en faisant glisser et déposer (les deux variables et leurs résultats).

Pour commencer le processus de correspondance à partir de zéro, on utilise le bouton **« reset »** réinitialiser, ce qui entraînera la correspondance suivante (Fig5.10) :

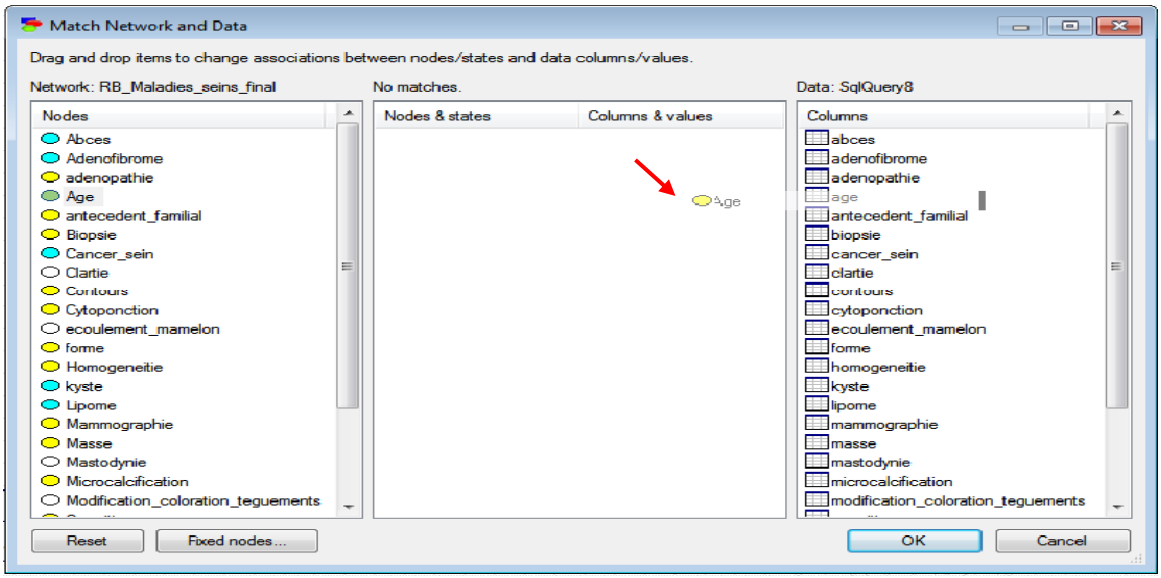


Fig.5. 10 Le processus de drag-and-drop

En faisant glisser et déposer de gauche (nœud) vers la droite (donnée) cela mettra la variable Age venu des deux listes au milieu (Fig.5.11) ; ce qui veut dire que la correspondance est effectuée.

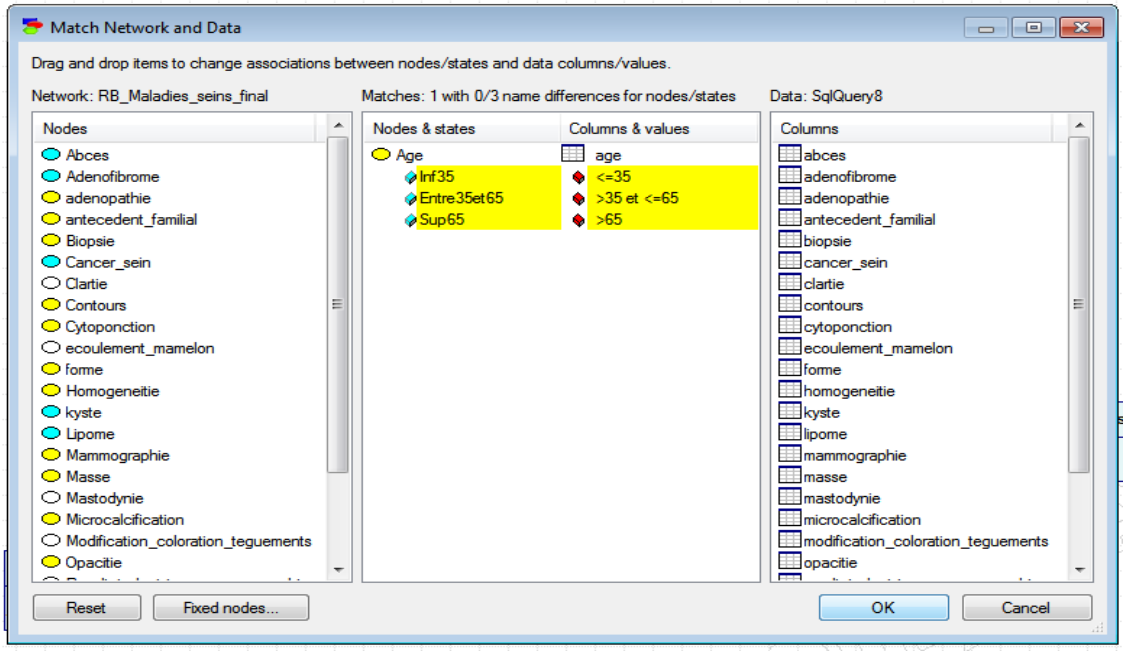


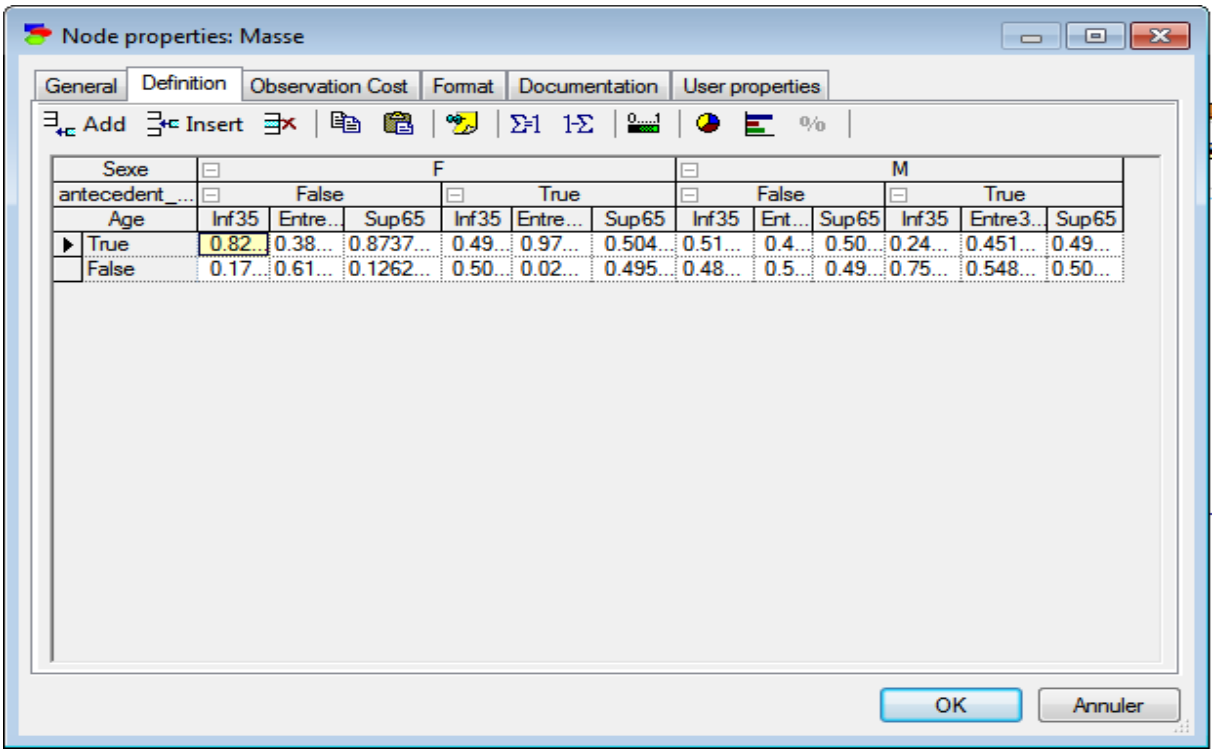
Fig.5. 11 Affectation de la correspondance pour la disparité entre le réseau et les données

Après avoir éliminé la disparité par la correspondance manuelle, on clique sur ok le réseau bayésien sera affiché avec la distribution de probabilités représenté à l'aide des tables de probabilités conditionnelles de chaque variable de réseau.

Chaque ligne d'une table de probabilités conditionnelles doit être égale à 1 car ils représentent un ensemble exhaustif de cas pour une variable. Pour une variable booléennes X_i, une fois qu'on sait que la probabilité qu'une valeur soit vraie est *p*, la probabilité qu'elle soit fausse doit être 1- *p.* En général, la table pour une variable booléenne ayant *k* parents booléens contient 2^k rangés ; si le nombre maximal de parents est *k,* alors le réseau demande O ($n2^K$) nombres (complexité linéaire en n), mais exponentiel en *k* pour la table complètes.

Deux cas se présentent selon la position d'une variable Xi dans le réseau bayésien :

- La variable Xi n'a pas de variable parente : on doit préciser la loi de probabilité marginale de Xi.
- La variable Xi possède des variables parentes : on doit exprimer la dépendance de Xi en fonction des variables parentes, au moyen de probabilités conditionnelles, la (Fig.5.12) présente une distribution de probabilités pour le nœud « masse » dans l'onglet **définition** de la fenêtre **node properties**.



Sexe	F						M					
antecedent_...	False			True			False			True		
Age	Inf35	Entre...	Sup65	Inf35	Entre...	Sup65	Inf35	Ent...	Sup65	Inf35	Entre3...	Sup65
True	0.82...	0.38...	0.8737...	0.49...	0.97...	0.504...	0.51...	0.4...	0.50...	0.24...	0.451...	0.49...
False	0.17...	0.61...	0.1262...	0.50...	0.02...	0.495...	0.48...	0.5...	0.49...	0.75...	0.548...	0.50...

Fig.5. 12 Table de probabilités conditionnelles pour la variable "Masse" liée aux trois parents

5.5.4. Algorithme d'inférence

L'utilisation d'un algorithme d'inférence permet d'interroger le modèle par une propagation de toute probabilité à priori sur la probabilité des autres nœuds, on obtient un nouveau tableau des probabilités sur chaque nœud, une sorte de nouvel état des lieux. Dans notre cas, on vient de réaliser un modèle de comportement probabiliste par les réseaux bayésiens sur le système de diagnostic SADMseins.

Pour calculer l'inférence dans notre cas, nous avons opté pour le moteur de calcule d'inférence basé sur l'algorithme de jonction qui passe par deux phases :

1) La phase de construction : c'est la transformation du graphe initial en un arbre de jonction. Cette transformation se fait par les étapes suivantes :

A. Moralisation du graphe (Fig.5.14)
B. Triangulation du graphe (Fig.5.15)
C. Création d'un arbre appelé arbre de jonction de jonction : il est constitué de 301 cliques.

2) La phase de propagation : il s'agit de la phase de calcul probabiliste. La phase de propagation dans l'algorithme de l'arbre de jonction « *Jonction Tree* » qui est basé sur la propagation de messages (*message passing*) est difficile à manipuler pour les grands réseaux, pour cela nous avons utilisé le logiciel **GeNIe** qui permet l'apprentissage automatique des tables de probabilités à partir des données.

Pourquoi utiliser GeNIe :

- GeNie, permet de prendre dès aujourd'hui les bonnes décisions de demain ;
- Il est facile à exploiter ;
- Evolutif.

Il met en œuvre une variété d'algorithmes. L'utilisateur peut choisir quel algorithme doit être utilisé pour la mise à jour en sélectionnant un algorithme dans la liste (Fig.5.13) sachant que l'algorithme de clustering (arbre de jonction) est l'algorithme par défaut de GeNIe et devrait être suffisant pour la plupart des applications.

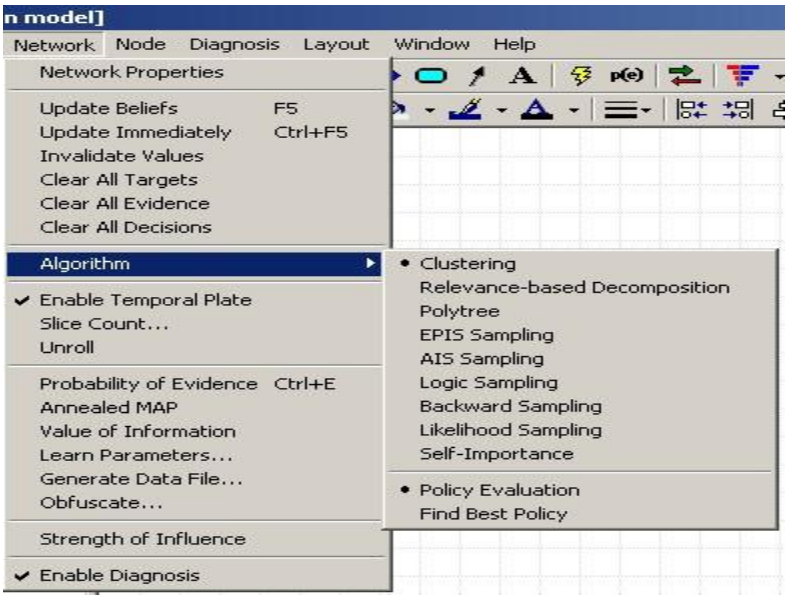


Fig.5. 13 La liste des choix des algorithmes d'inférence

Etape de moralisation

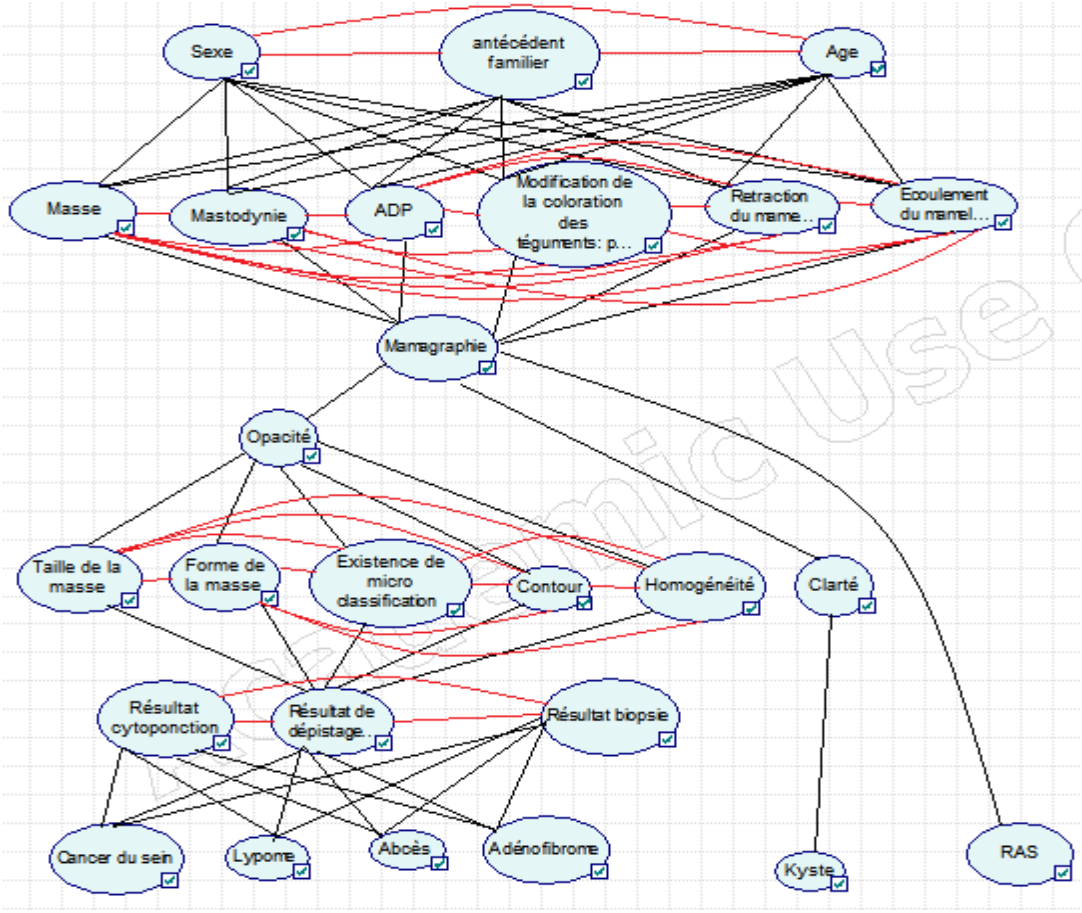


Fig.5. 14 Graphe moralisé

Etape de triangulation

Selon l'algorithme de Kjærulff [kja 90] l'ordre d'élimination est : {RAS, kyste, clarté, adénofibrome, abcès, lipome, cancer du sein, résultat dépistage-mammo, mammographie, cytoponction, biopsie, taille, forme, microcalcification, contour, homogénéité, opacité, masse, Mastodynie, ADP, modification CT, retraction mamelon, écoulement mamelon, âge, antécédent f, sexe}, les arêtes à ajouter sont montrées dans la (Fig.5.15) dont la couleur verte

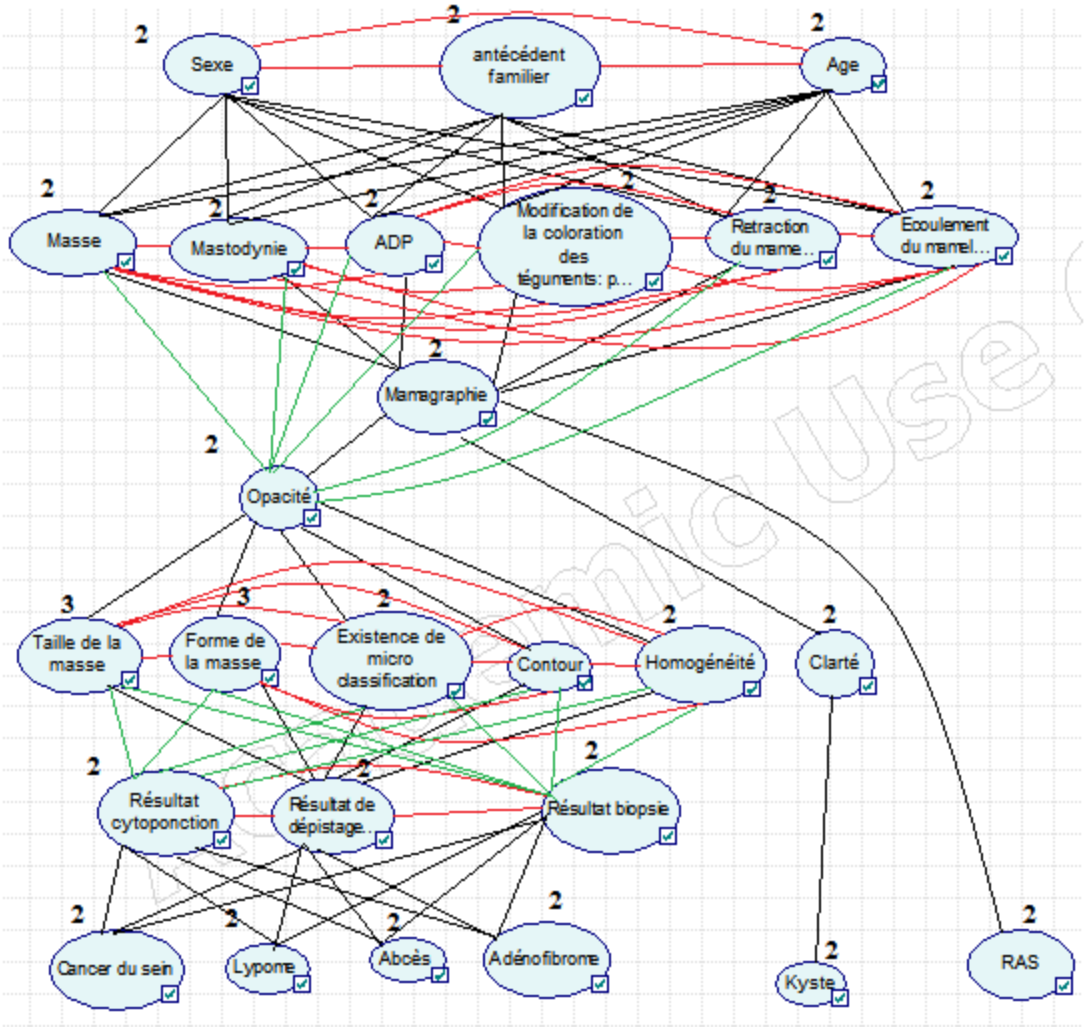


Fig.5. 15 Etapes de triangulation selon l'algorithme de Kjærulff

5.5.3.1. Pourquoi choisir l'algorithme arbre de jonction

L'arbre de jonction (clustering ou clique-tree propagation algorithm) a été introduite par Jensen, Lauritzen & Olesen est un algorithme d'inférence exacte dans les réseaux bayésiens, il est efficace pour le calcul de la distribution marginale d'une variable.

. Il est plus général ce qui concerne la structure graphique. Comme dans notre cas le choix était l'utilisation de l'algorithme de l'arbre de jonction vue que :

- Notre modèle graphique est plus général comporte plusieurs racine (plus de deux parents), l'utilisation de cet algorithme s'adapte mieux à la structure.
- Raisonnement avec données complètes d'où le principe de l'inférence exacte de l'arbre de jonction
- Algorithme arbre de jonction est un algorithme exact qui permet de parcourir tous l'espace de recherche de notre graphe.
- La précision du calcul de la loi marginale d'une variable ou de sa loi conditionnelle relativement à un ensemble d'observations.

Ce type d'inférence appliqué dans l'algorithme d'arbre de jonction, appelé "***mise à jour***" des probabilités est essentiel en particulier dans les applications à l'élaboration de ***diagnostic***, ou l'on doit reconsidérer la situation en fonction d'une ou plusieurs nouvelles observations. La complexité de cet algorithme au moment de la propagation des messages est de $O\sum_{i=1}^{N_C} n_e\ (ci)$ ou N_c est le nombre de cliques de l'arbre de jonction et $n_e(c_i)$ est le nombre d'états de cliques C_i. Ainsi, pour réduire cette complexité il est nécessaire de construire des cliques ayant un petit nombre de variable (et c'est possible avec des variables ayant un petit nombre d'état).

S'agissant en général, de variables aléatoires à nombre fini de valeurs, ***il s'agit ici essentiellement d'un problème d'optimisation de calcul, puisque celui-ci devient de plus en plus lourd suivant la complexité du graphe relativement à la fois au nombre de variables et au nombre de valeurs prises par ces variables.*** Cependant ces problèmes restent relativement complexes et donnent lieu à de nombreuses recherches.

Malgré la diversité des algorithmes d'inférence qui ont tous l'objectif de ***diminuer la complexité de temps et de calcul.*** L'inférence probabiliste sur les réseaux bayésiens restent une tache intense et difficile pour tous les algorithmes qui porte des avantages et inconvénients mais ils ont un point commun qui est la complexité polynomiale. [COO87] montre que ce problème est NP-difficile, quoique plusieurs recherches aient été dirigées pour développer des algorithmes efficaces pour ce genre de problèmes.

5.5.5. La validation du système SADMseins (Résultats)

Pour valider l'outil, nous avons implémenté un modèle avec 50 liens et 26 nœuds, comme le montre la (Fig.5.18) ; le système résolu représente une modélisation raisonnable du processus de diagnostic de cinq maladies mammaires validées par un expert. Le (Tableau.5.1) ci-dessous illustre les différents nœuds du réseau bayésien chacun avec son identificateur Id et les valeurs des variables :

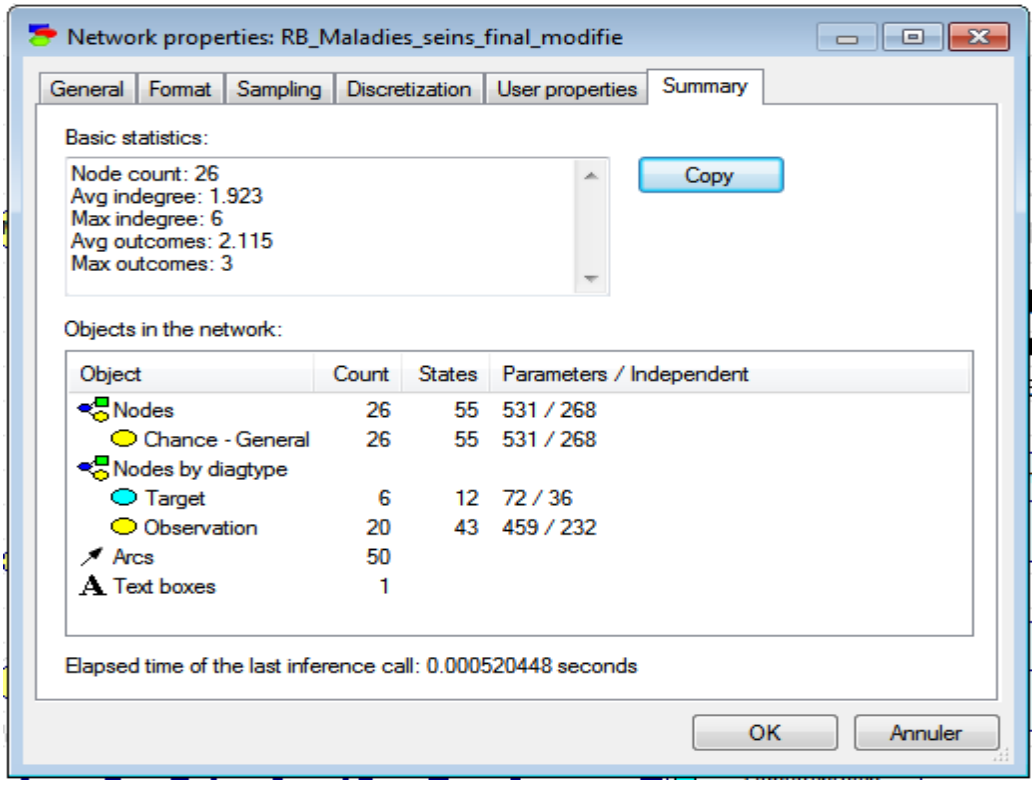


Fig.5. 16 Fenêtre de propriétés graphiques du réseau bayésien

La fenêtre **network properties** (Fig.5.16) décrit les statistiques qui se concentrent sur les propriétés structurelles du réseau, tels que le nombre de nœuds de chaque type dans le réseau, la moyenne et le maximum en degré (le nombre de parents d'un nœud), le nombre moyen et le nombre maximal de résultats des nœuds, le nombre de nœuds par leur type de diagnostic, le nombre d'arcs et le nombre de zones de texte, et enfin le nombre d'états et de paramètres.

SADMseins présenté contient 26 nœuds, dont 6 sont des nœuds cible de diagnostiques (maladies) et 20 sont des nœuds d'observation. Le nombre d'arcs (50) et le niveau moyen en degré (1.923) donnent une idée de la complexité structurelle du réseau.

Id du nœud	Valeur
Antécédent familiaux	Oui, non
Age	{< 35}, {entre 35 et 65},>65
Sexe	Femme, homme
Modification de la coloration des téguments	Oui, non
Adénopathie	Oui, non
Mastodynie	Oui, non
Rétraction du mamelon	Oui, non
Ecoulement du mamelon	Oui, non
Masse	Oui, non
Mammographie	Oui, non
Opacité	Oui, non
Clarté	Oui, non
Taille de la masse	<2mm, 2-5mm,>5mm
Forme de la masse	Round, spéculé, ovale
Microcalcification.	Oui, non
Homogénéité	Homogène, hétérogène
Contour	Régulière, irrégulière
Biopsie	Bénin, malin
Cytologie	Bénin, malin
Résultat de dépistage de la mammographie	Bénin, malin
Cancer des seins	..%
Adénome fibrome	..%
Abcès	..%
Kyste	..%
Lipome	..%
Rien à signalé	..%

Tableau.5. 1 Les variables du réseau bayésien

La base de données comporte l'ensemble des tables de probabilités à priori et conditionnelles. Notre réseau bayésien contient 26 tables au totale.

L'utilisation du RB repose sur l'algorithme d'inférence qui calcule la probabilité de chaque variable à partir d'un ensemble de valeurs fixées à priori.

GeNie permet à partir du réseau bayésien de faire entrer les évidences (les valeurs) pour l'ensemble des signes cliniques, biologique et l'attribut d'imagerie médicale caractérisant le

cas traiter(les signes qui apparaissent sur le patient). La (Fig.5.17) montre l'instanciation d'une variable par une évidence.

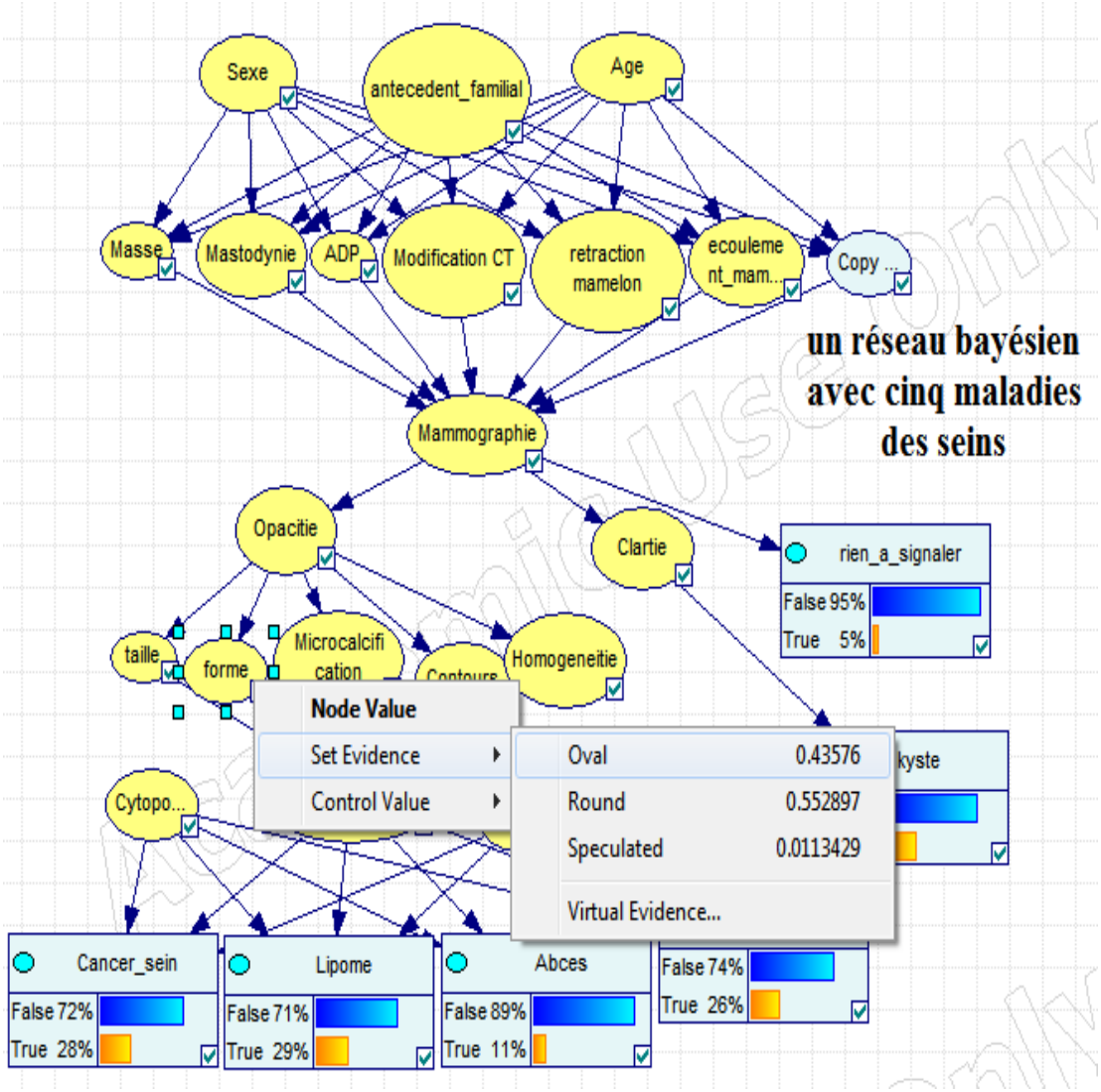


Fig.5. 17 l'instanciation des variables

Après entrée les informations de chaque variable du réseau bayésien, l'état des nœuds qui sont sélectionnés pour établir un diagnostic prennent l'état évidence décrite par le symbole (⊟) au lieu de l'état valide ☑ comme le montre la (Fig.5.18)

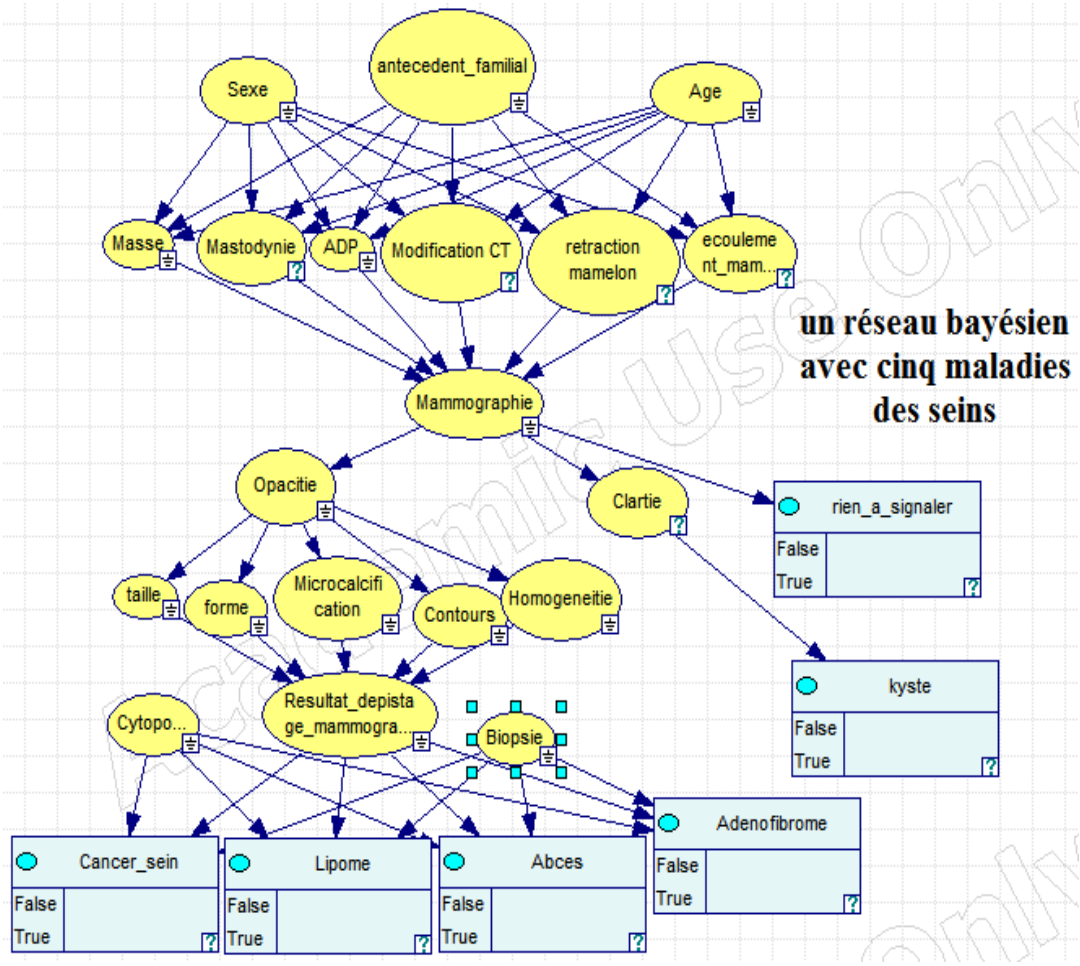


Fig.5. 18 Réseau bayésien en état actif

On procède alors à une première inférence avec le logiciel GeNie, on peut présenter l'ensemble actuel des évidences prisent pour obtenir un diagnostic dans le (Tableau.5.2) comme exemple d'exécution.

Sexe	Age	Antécédent-Familiaux	Masse	Adénopathie	Opacité	Taille
f	>65	oui	oui	oui	oui	<2cm
Forme	Homogénéité	Contour	Micro calcification	Biopsie	Cytoponction	Résultat-mammographie
round	hétérogène	irrégulière	oui	maligne	maligne	bénigne

Tableau .5. 2 Exemple d'activation du système cas de tumeur maligne

L'algorithme de clustering transforme le graphe en un arbre de jonction ou l'inférence dans les réseaux bayésiens se fait par le passage de messages entre cliques.

GeNIe cherche la variable du dernier niveau ayant la probabilité maximale sachant qu'elle sera le résultat du diagnostic.

Une fois les extensions de diagnostic activées, nous cliquons sur le bouton Test Diagnostic ( ) de la barre d'outils cela met à jour les distributions de probabilité à la lumière des preuves observées (évidences). Une fois que la fenêtre de test est ouverte, elle apparaîtra comme indiqué ci-dessous (Fig.5.19) :

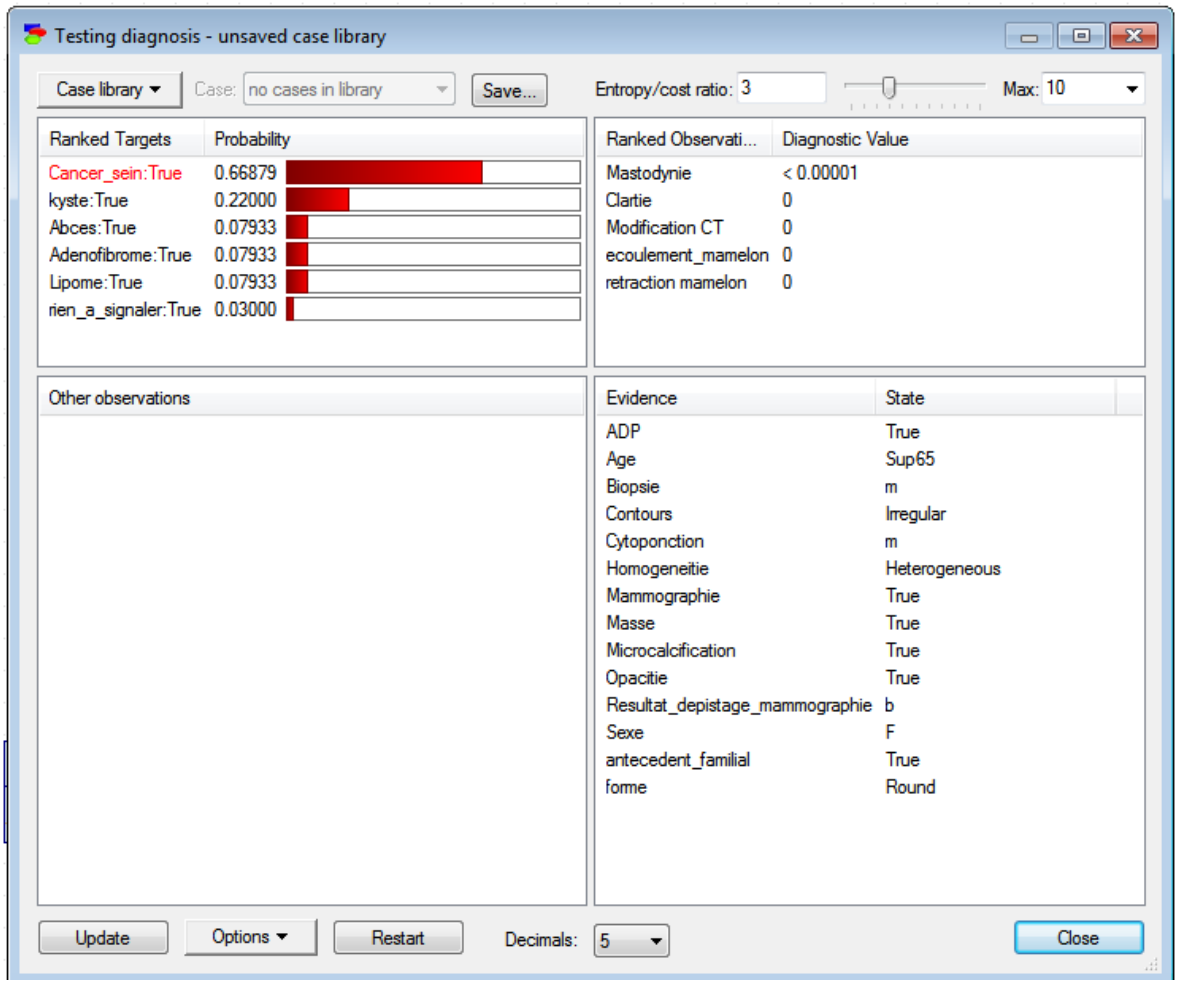


Fig.5. 19 Interface de diagnostic fourni par GeNIe pour la maladie du cancer du seins (dans les cas de tumeur maligne)

Le système SADMseins implémenté sous GeNIe est équipé d'une simple interface utilisateur dédiée qui nous a permet d'entrer dans diverses observations telles que les symptômes et les résultats des tests médicaux et affiche la distribution de probabilité sur diverses maladies les plus fréquentes des seins possibles dans l'ordre de la plupart des moins

susceptibles. Il classe également les observations possibles en fonction de leur valeur de diagnostic.

La fenêtre de diagnostic comporte quatre volets, qui peuvent être décrits comme suit :

- Le volet supérieur gauche concerne les nœuds cibles, qui sont des nœuds non observables et classés, représentant généralement différentes maladies.
- Le volet supérieur droit concerne les observations classées, qui sont des nœuds observables qui n'ont pas encore été observés (non pas étaient pris lors de la validation).
- Le volet inférieur gauche répertorie d'autres observations, qui sont les observations qui ont été désignées comme obligatoires et ne sont pas encore observées. Leur désignation comme obligatoire indique qu'ils sont faciles à observer ou font autrement partie des premières étapes et doivent d'abord être observés. Ils sont affichés en caractères rouges pour attirer l'attention de l'utilisateur.
- Le volet inférieur droit contient tous ces nœuds parmi les observations classées qui ont été observés (les évidences).

On procède une deuxième inférence avec le logiciel GeNIe, on fait entrer l'ensemble des évidences prisent pour obtenir un diagnostic dans cas bénin le (Tableau.5.3) comme exemple d'exécution.

Sexe	Age	Antécédent-Familiaux	Masse	Adénopathie	Opacité	Taille
f	< 35	non	oui	oui	non	>5cm
Forme	Homogénéité	Contour	Micro calcification	Biopsie	Cytoponction	Résultat-mammographie
round	homogène	irrégulière	oui	bénigne	bénigne	bénigne

Tableau .5. 3 Exemple d'activation du système cas de tumeur bénigne

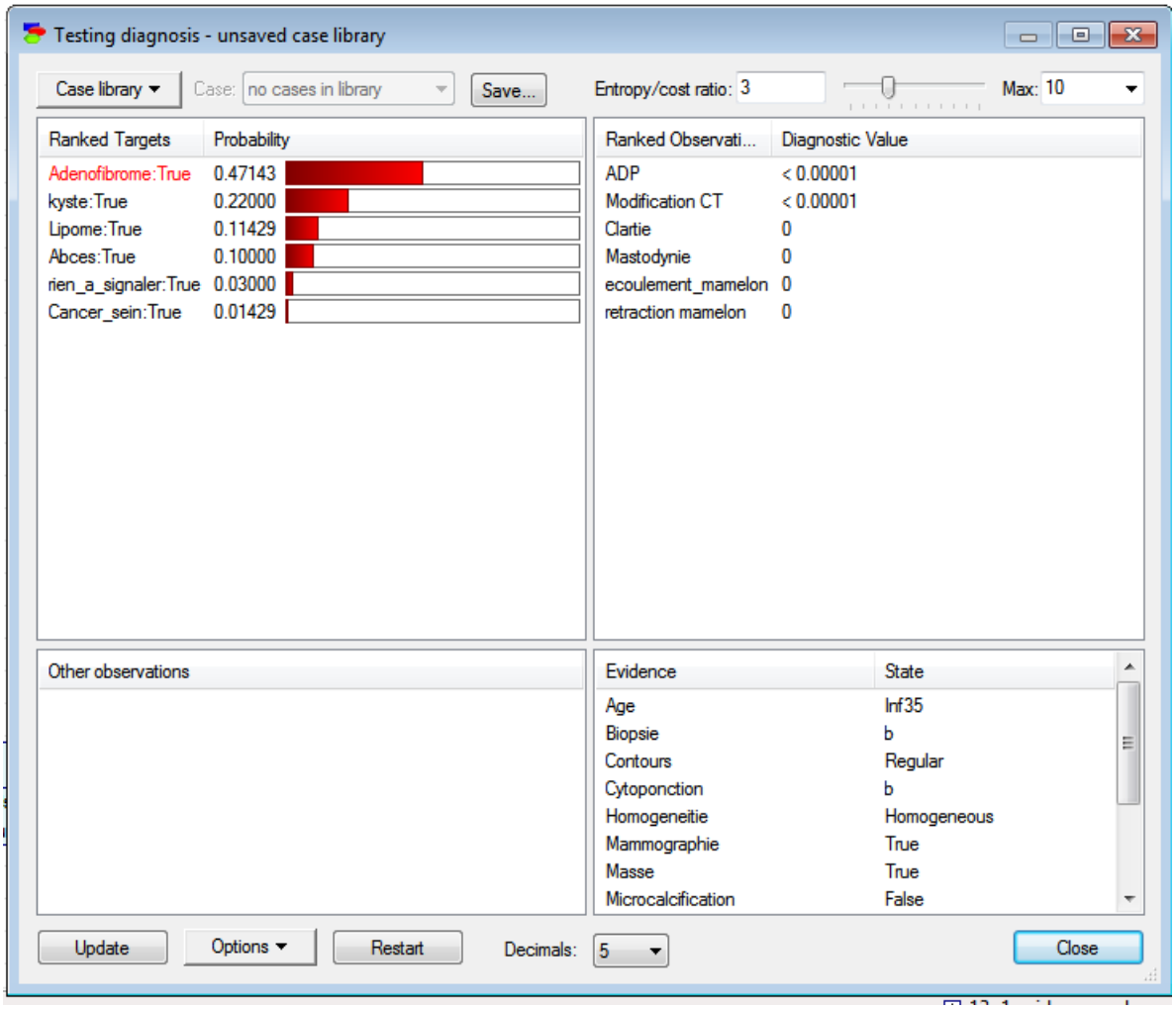


Fig.5. 20 Interface de diagnostic pour la maladie adénomefibrome (dans les cas de tumeur bénigne)

5.6. Discussion des résultats de diagnostic de notre système

La sélection des attributs pour chaque niveau du réseau bayésien pour les deux aspects : dans le cas de la maladie maligne et le cas de la maladie bénigne. Les résultats obtenus sur les (Fig.5.19) et (Fig.5.20) montrent que le système peut fournir la meilleure action du raisonnement de diagnostique aux médecins.

GeNIe fournit également une interface pour effectuer un diagnostic. Une fois qu'un modèle de diagnostic est créé, le modèle peut être chargé dans GeNIe et utilisé pour appliquer le diagnostic. Après avoir établi des éléments de preuve pour certaines variables contextuelles et probantes, GeNIe est capable de calculer la ou les maladies les plus probantes étant donné - cette évidence.

Il permet également de présenter les résultats dans un graphique les probabilités de toutes les maladies et aussi les probabilités de chaque maladie par rapport à toutes les maladies étudiées. Les résultats obtenus dans les deux diagrammes montrent que tous les observations (symptômes) dominent au cours des résultats, et le taux de maladies bénignes selon les essais est plus fréquent que la maladie maligne. Le système est exactement le principe de l'algorithme de clustering.

5.7. Les résultats apportés par la maintenance

Dans cette section nous allons présenter un exemple montrant l'utilité de la maintenance qui porte un bénéfice pour notre réseau bayésien et se résume dans l'analyse des résultats, dans l'évaluation des couts de maintenance et par la notion de mise à jour porté par la maintenance qui rend notre logiciel plus fiable.

L'espace de travail de GeNIe comporte divers éléments comme la barre d'outils standard qui permet d'ajouter ou de supprimer des nœuds et des arcs, ce dernier facilite la tâche de maintenance pour les experts. À ce niveau, nous identifions les actions de maintenance apportées avec leurs probabilités a priori et conditionnelles en sachant que leurs types sont discrets, et sont les suivantes : (A : désigner une action de maintenance)

- **A1** : Prendre des pilules.
- **A2** : Obésité.

Nous verrons leur influence sur la maladie concernée (cancer du sein). Une fois que cette variable importante est identifiée, l'aide au diagnostic prend part à l'intégration d'une action de maintenance sur cette variable considérée comme importante. Pour cela, nous considérons les tâches de maintenance comme de nouveaux nœuds du réseau bayésien apportés avec une probabilité à priori.

Le nouveau réseau bayésien avec action de maintenance est donné à la (Fig.5.21). Ces analyses ont été faites par deux actions de maintenance, mais il est possible de combiner deux actions ou plus.

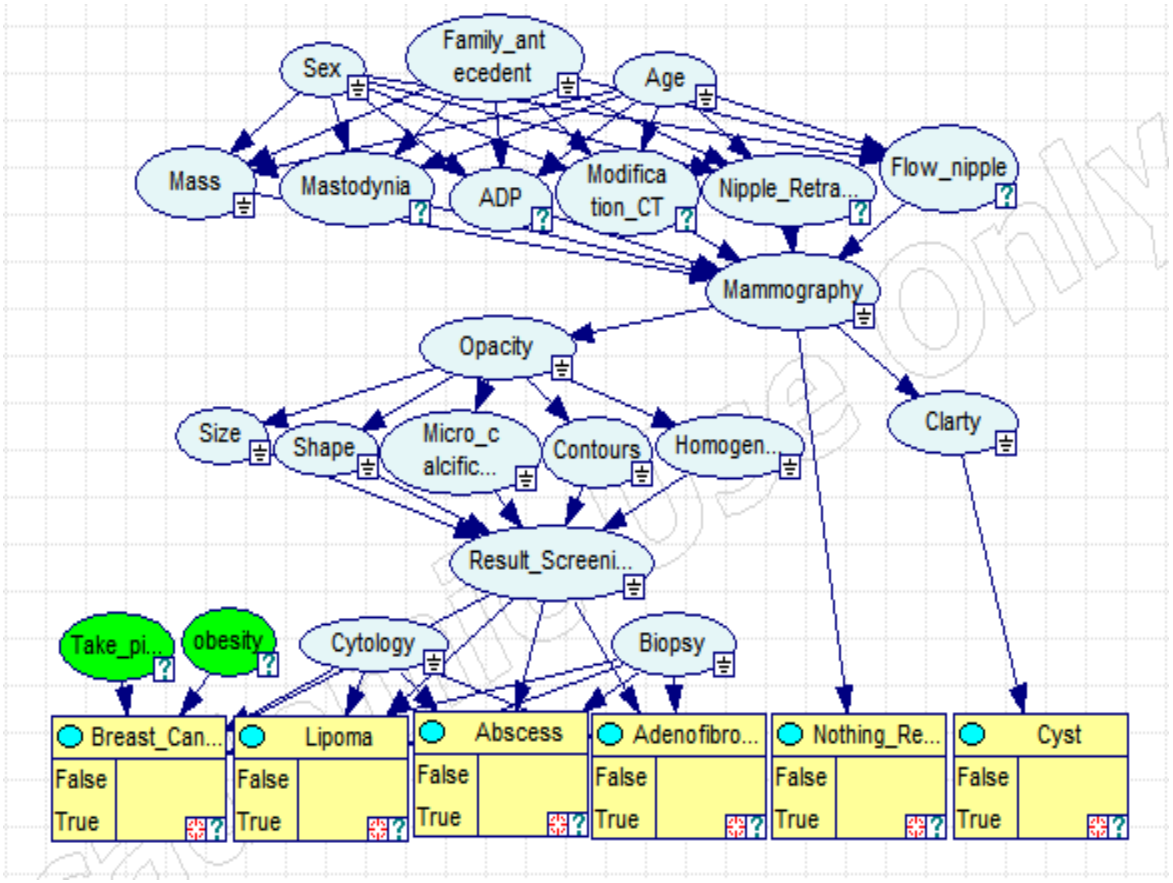


Fig.5. 21 Diagramme général du système de modélisation avec deux actions de maintenance

L'utilisateur (médecin) sélectionne à tous les niveaux, les évidences de la variable du réseau bayésien qui apparaît sur les patients. Le choix des tâches de maintenance est effectué par le médecin en fonction de l'état du patient. Le système donne la possibilité d'ajouter un nouveau nœud accompagné de l'ajout des arcs et le médecin doit décider de quelle variable du réseau bayésien les actions vont agir dont le remplissage des tableaux de probabilité de l'action de maintenance en sachant que cette dernière est toujours parente. Par conséquent, un nouveau réseau bayésien est obtenu temporairement.

Le système lance le processus d'activation et les variables sont initialisées par un ensemble de probabilité. Une nouvelle inférence exacte est effectuée en tenant compte de l'action de la maintenance. Nous implémentons une recherche en profondeur par passage de message en calculant la distribution conjointe qui représente la probabilité de toutes les combinaisons.

Les variables doivent être classées par ordre topologique, c'est-à-dire qu'une variable ne peut pas précéder ses parents dans le réseau. Pour calculer les probabilités d'une variable, les valeurs de ses parents doivent déjà être connues. L'ajout d'action (ou plus) de maintenance découle **de l'Algorithme 1** étant donné que l'action de maintenance peut être supprimée à tout moment et que sa valeur peut également être modifiée en conservant la structure fixe, les résultats de la maintenance sont représentés dans la (Fig.5.22).

Algorithme 1. Algorithme d'ajout des actions de maintenance dans le réseau bayésien.

-Choisir un nombre des actions à ajouter A1, ..., An
For i=1 to n **do**
Ajouter Ai dans le réseau fixé
For j=1 to n **do**
-Sélectionner ces feuilles dans le réseau bayésien Xs, ..., Xn tel que
P (Xj) = P (Xj\ parent (Ai)).
End for j
End for i

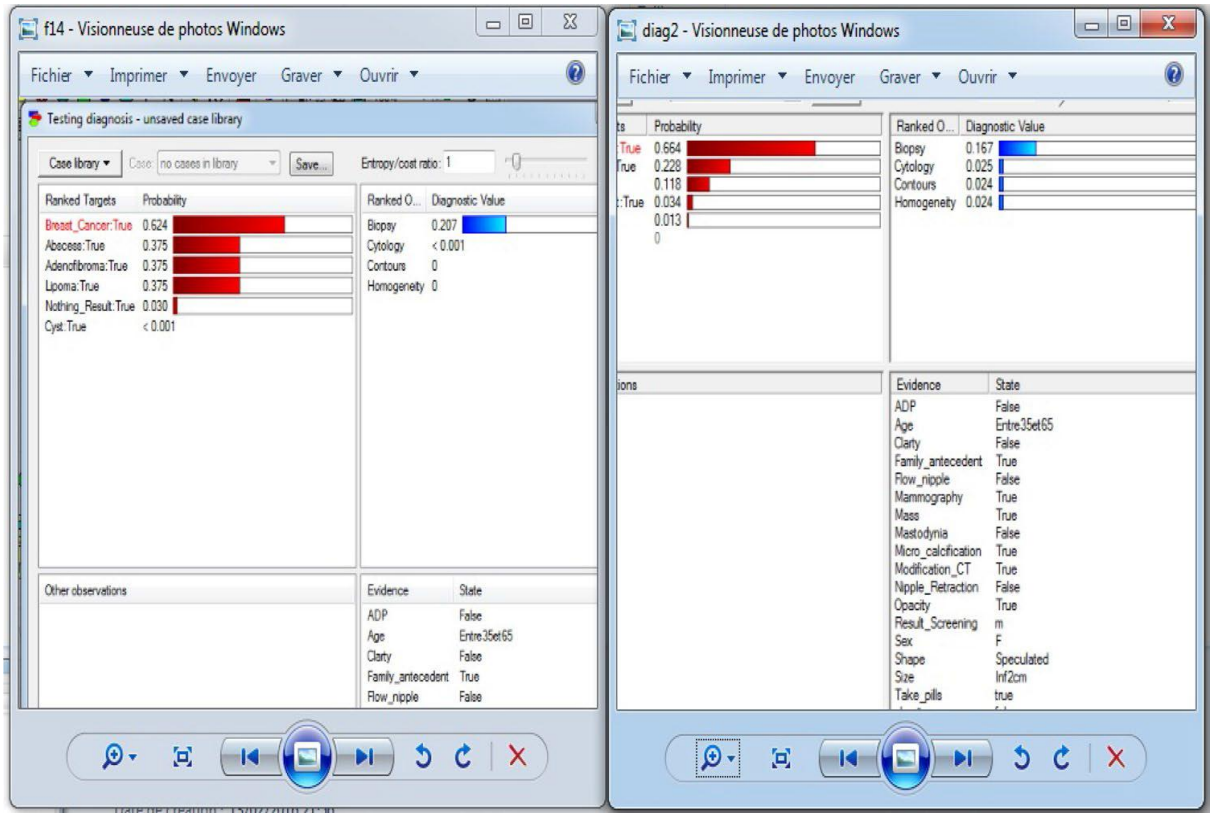


Fig.5. 22 Interface de diagnostic pour les résultats avant et après l'ajout des actions de maintenance

Les résultats apportent une augmentation de la probabilité d'apparition du cancer du sein par rapport au résultat ((tableau.5.4), (Fig.5. 19) et (Fig.5. 22)). Notre intérêt à ce niveau est basé sur les changements apportés avant la maintenance et après avoir ajouté d'autres informations et essentiellement un problème d'optimisation du calcul, puisqu'il devient de plus en plus lourd en raison de la complexité du graphe par rapport au nombre de variables et au nombre de valeurs prises par ces variables. Dans notre cas, les actions ajoutées portent des valeurs binaires. Nous avons choisi, dans notre exemple, deux actions de maintenance "obésité" et "prendre de pilules". La validation a apporté des résultats appropriés par rapport aux opinions des experts qui considèrent les attributs ajoutés dans notre étude comme des facteurs primordiaux qui modifient la probabilité vers une amélioration avancée. L'addition a augmenté la probabilité d'avoir un cancer du sein avec un taux de 4% décrit dans le tableau ci-dessus.

L'ajout des actions			
Obésité		**Prendre de pilules**	
Valeur	**Probabilité**	**Valeur**	**Probabilité**
Oui	0,4	**Oui**	0,4
Non	0,6	**Non**	0,7

Tableau.5. 4 Table de probabilité des actions de maintenance

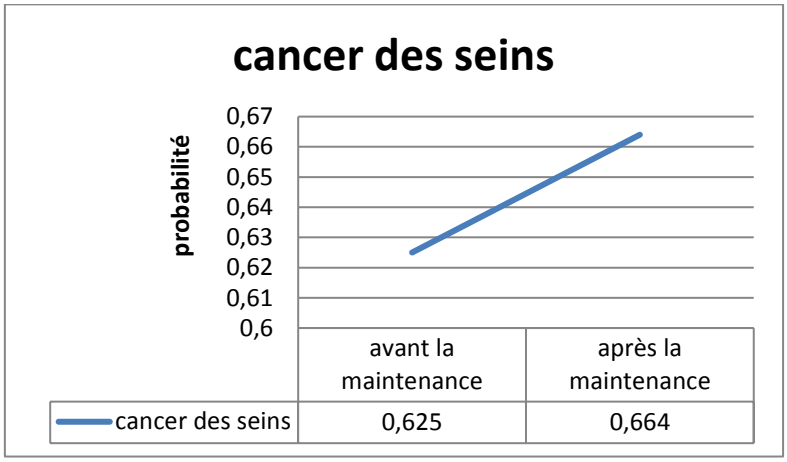


Fig.5. 23 Diagramme général modélisant les résultats avant et après l'ajout des actions de maintenance

Cela permet de dire que les actions de maintenance dans ce cas aident à réduire le taux d'erreur en augmentant la probabilité pour la maladie cancer du sein et qui donne un bon diagnostic tout en s'éloignant de l'incertitude.

5.8. Évaluation de la performance

Il y a toujours une variabilité biologique entre les personnes, et cela comprend les signes et les symptômes que présentent les personnes quand elles sont malades. Ainsi, aucun système de diagnostic n'est parfaitement précis : les personnes peuvent varier dans leur degré d'atteinte par une maladie, même si elles ont les mêmes résultats au test.

Il est évident que nous voulons que le système de diagnostic soit précis : ils ne devraient ni manquer des cas, ni classifier des personnes en santé comme des personnes malades. Dans ce contexte nous avons vérifié les critères d'évaluation concernant les performances de diagnostic du système SADMseins en termes de sensibilité, spécificité et précision à partir des résultats présentés à la (Fig.5.24).

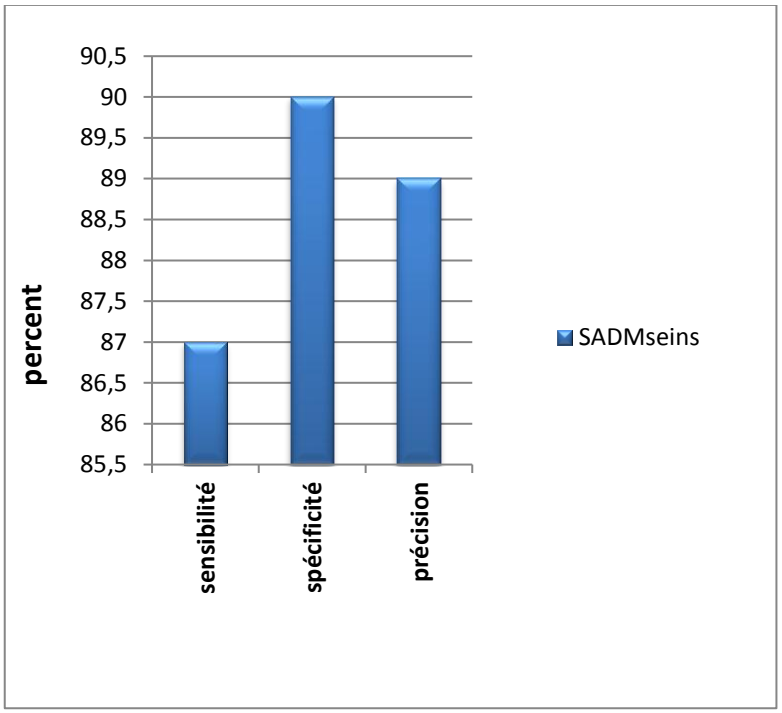


Fig.5. 24 Histogramme des indices statistiques obtenus pour le SADMseins (n = 100)

La sortie du système de diagnostic médical SADMseins sur l'échantillon de validation (n = 100) a montré une classification correcte chez 26 des 30 patients atteints de cancers du sein (dans les cas de tumeurs malignes) et chez 63 des 70 patients atteints de tumeurs bénignes. Les résultats obtenus de la sensibilité, de la spécificité et de la précision sont respectivement de 87%, 90% et 89%. Les résultats du système proposé sont très encourageants et nous permettent de tester en toute confiance le système malgré l'absence d'un nombre beaucoup plus grand de cas de test requis pour évaluer de manière exhaustive le système.

5.9. Conclusion

Nous avons présenté dans ce chapitre un nouveau modèle de réseau bayésien pour le diagnostic des maladies les plus fréquentes des seins. Nous avons mis en place un système d'aide au diagnostic médical pour résoudre les problèmes de décision avec des données complètes en essayant de fournir aux médecins un outil de raisonnement informatique pour améliorer leurs décisions. Le système agit comme le raisonnement des experts grâce aux réseaux bayésiens qui présentent une méthodologie de la représentation du savoir pour exprimer des relations entre les variables en utilisant la théorie des probabilités pour gérer le problème de l'incertitude médicale et faire une vision compréhensible pour un non-spécialiste.

Le présent travail montre brièvement que l'algorithme d'inférence bayésien trouvé dans la littérature peut être facilement implémenté et adapté pour répondre aux systèmes avec des restrictions de diagnostic, car l'efficacité de cet algorithme dépend toujours de la qualité de la triangulation.

En général, trouver une triangulation optimale pour les graphes reste un problème NP-difficile. Notre étude nous a permis d'établir la formalisation d'un système pour un diagnostic des maladies du sein basé sur le réseau bayésien enrichi par le concept de la maintenance qui apporte un avantage pour les réseaux bayésiens résumée dans l'analyse des résultats.

Conclusion générale et perspectives

Le travail réalisé dans le cadre de cette thèse, présente une étude de conception pour développer un système d'aide au diagnostic médical à partir d'un modèle de représentation de connaissances de type Réseau Bayésien.

Les réseaux bayésiens représentent dans ce contexte une bonne méthodologie de représentation des connaissances. Cette approche permet d'exprimer les relations entre les variables basées sur la théorie des probabilités et de raisonnement sous incertitude liée au domaine médicale afin d'apporter une vision compréhensible pour un non spécialiste.

Ce travail s'intéresse en premier lieu à développer un nouveau modèle graphique de réseau bayésien pour l'aide au diagnostic des pathologies les plus fréquentes des seins tout en retirant les connaissances les plus pertinentes. La structure du réseau bayésien obtenu est évaluée uniquement à partir d'avis d'experts, elle permet une représentation de connaissances qualitatives et quantitatives exprimant l'incertitude décomposée en quatre niveaux, niveau clinique, niveau biologique, niveau imagerie médicale et niveau diagnostic.

Par ailleurs, l'évaluation de ce modèle nécessite une base de données qui doit être suffisamment importante pour pouvoir répondre à l'ensemble des questions dans un SADDM, à ce niveau, nous nous sommes concentrés dans un premier temps à collecter le maximum de dossiers en se basant sur l'ensemble des données cliniques, radiologiques et biologiques décrites dans le réseau bayésien afin de déterminer la probabilité de malignité ou bénignité.

Le second axe de travail de cette thèse est celui de l'intégration de la structure proposée ainsi que la base de données appropriée au logiciel GeNIe pour une mise en œuvre pratique d'un système qui a pour rôle principale l'aide au diagnostic des maladies les plus fréquentes des seins « ***SADMseins*** ».

L'évaluation de données se fait d'abords par l'apprentissage automatique (statistique) pour cela nous avons utilisés l'approche maximum de vraisemblance (**MV**) qui estime la probabilité d'un évènement par sa fréquence d'apparition dans la BDD collectées en aboutissant à la fin une distribution de probabilités complète à l'aide des tables de probabilités conditionnelles de chaque variable de réseau.

L'utilisation du réseau bayésien repose sur l'algorithme d'inférence exacte, Les entrées de notre système sont l'ensemble des attributs cliniques, biologiques et les paramètres d'imagerie médicale qui représentent des variables discrètes apportées avec leurs probabilités. L'algorithme JLO (clustering) calcule donc la probabilité de chaque variable à partir d'un ensemble de valeurs fixées à priori pour obtenir les probabilités des variables de sorties qui sont les maladies des seins, cet algorithme est à la base d'une famille plus complète d'algorithmes permettant de faire de l'inférence.

Pour évaluer le système, nous avons utilisé une base de données de 20 tables de probabilités pour 26 nœuds. Les résultats montrent que le système ***SADMsein*** est capable de fournir un bon diagnostic lors de l'évaluation grâce à l'outil GeNie qui marque sa puissance pour le développement des réseaux bayésiens par la capacité d'apprentissage automatique, par la simplicité de la manipulation, la représentation visuelle du modèle graphique est un atout majeur pour communiquer les résultats.

Nous avons pu conclure que le modèle proposé a montré son efficacité et a permis, dans notre cas, de pouvoir effectuer un diagnostic et une mise à jour de ce diagnostic, à chaque fois que de nouvelles observations ont été disponibles.

En effet, chaque système informatique notamment les systèmes qui essayaient de simuler les raisonnements incertains des médecins par fois nécessite en phase de fonctionnement d'un côté d'être maintenu pour garantir la qualité, la performance, et la compétence de ce système, et d'un autre coté pour contrôler, réagir face aux changements dans la dynamique du domaine avec ces différentes sources de connaissances et en raison d'erreurs dans le modèle construit.

Donc nous avons décri la maintenance bayésienne qui consiste à définir des variables supplémentaires notées comme des actions de maintenances pour mettre à jour le réseau bayésien et améliorer les performances du système SADMsein.

Les perspectives du présent travail se situent dans au moins les directions suivantes :

- Proposition de plusieurs modèles de structures bayésien de plusieurs experts pour le diagnostic des maladies des seins et la combinaison de ces modèles en une seule structure.
- Le deuxième problème était la difficulté d'obtenir des données complètes de notre réseau par des études statistiques afin d'obtenir toutes les preuves. Dans les applications pratiques, les bases de données 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.

- La prise en compte des coûts de nouvelles variables (actions de maintenance) dans le cadre d'optimiser les politiques de maintenance ; ces coûts pourront par exemple pondérer sur les arcs du graphe.
- Un autre aspect important est ce que nous appelons le phénomène de « 1'explosion » des tables de probabilités. Lorsqu'un nœud X a plusieurs causes probables (donc plusieurs parents) et qu'en plus, chacun des nœuds représentants les causes à plusieurs états possibles, la table de probabilités associée au nœud X peut devenir extrêmement complexe et difficile à gérer. Si, en plus, il n'y a pas suffisamment de données pour initialiser nos tables de probabilités (plus les tables sont grosses et complexes, plus nous avons besoin d'un grand échantillonnage).
- Établir une étude comparative pour décrire la pertinence a priori du modèle construit.
- Enfin, nous comptons élargir la base de données avec plus de cas possible afin d'obtenir un diagnostic le plus précis possibles et avec la plus grande certitude et fournir au médecin les outils pour la meilleure action thérapeutique pour réguler l'état de santé du patient.

Références bibliographiques

[ALE 10] Aleksovska S. L., Loskovska S., Clinical decision support systems: medical knowledge acquisition and representation methods, IEEE International Conference on Electro/Information Technology (EIT), pp. 1-6, 2010.

[ALS 12] Alsun M. H., Indexation guidée par les connaissances en Imagerie médicale, Thèse de doctorat, Ecole Doctorale-sicma, L'Université européenne de Bretagne, Janvier 2012.

[BAY 91] Bayes.T, An essay towards solving a problem in the doctrine of chances (1763), 189215, Bayesian Statistics: Principles, Models, and Applications, 1991.

[BEC 99] Becker A., Naim P., Les réseaux bayésiens, EYROLLES, première Edition, 1999.

[BEL02] Bellot D., Fusion de données avec des réseaux bayésiens pour la modélisation des systèmes dynamiques et son application en télémédecine, thèse doctorat, Université de Bruxelles, 2002.

[BEN 06] Benoit La voie, Apprentissage bayésien-Synthèse de lectures, (Séminaire sur l'apprentissage automatique), 2006.

[BEN 08a] Benferhat Salem, Tayeb Kenaza, Aïcha Mokhtari,Tree- Augmented Naïve Bayes for Alert Correlation, Dans 3rd conference on Advances in Computer Security and Forensics (ACSF'08), pp.45–52, jul 2008.

[BEN 08b] Benferhat Salem et Karim Tabia, Novel and anomalous behavior detection using Bayesian network classifiers, Dans International Conference on Security and Cryptography (SECRYPT'08), pp.13–20,2008.

[BEN 10] Benoit, Monique. Imagerie médicale, corps des femmes et regard occidental : une analyse de l'incertitude médicale autour du cancer du sein, Canadian Woman Studies, 2010, vol. 28, no 2.

[BEN 04] Ben Amor Nahla, Salem Benferhat, Zied Elouedi, Réseaux Bayésiens Naifs et Arbres de Décision dans les Systèmes de Détection d'Intrusions, Dans quatorzième Congrés Francophone AFRIF AFIA sur la Reconnaissance des Formes et l'Intelligence Artificielle (RFIA'04), pp.1175–1184, Toulouse France, jan 2004.

[BOU 05] Boucher A., Réseaux Bayésiens, Rapport du travail d'Intérêt personnel Encadré, Semestre II, pp. 1-29, 2005.

[BOU 12] Bouaziz M.F., Zamai E., Hubac S., Modélisation de l'état de santé d'un équipement de fabrication par une méthode probabiliste : Application aux ateliers semi-conducteurs, 9e Conférence Internationale de Modélisation, Optimisation et Simulation - MOSIM'12- Bordeaux – France, 2012.

[BOU 13] Bouaud J., Falcoff H., Séroussi B., Simultaneously authoring and modeling clinical practice guidelines: a case study in the therapeutic management of type 2 diabetes in France, Studies in Health Technology and Informatics, 186: pp.108–112, 2013.

[BRA 16] Brahim Ait Skourt, Rapport projet de fin d'études psychiatrie de liaison, faculté des sciences et techniques Fès, 2016

[BRO 05] Brossier J.M., Une introduction aux modèles graphiques, Laboratoire des images et des signaux, Février 2005.

[CAN 04] CANEL Christophen, Réseaux Bayésiens, Ecole d'ingénieursde Genéve, Mémoire PFE, 2004.

[CAS 96] Castillo E., Gutiérrez J. M., Hadi A. S., A new method for efficient symbolic propagation in discrete Bayesian networks, Networks, 28(1), pp.31-43, 1996.

[CHE 00] Chesnevar C.I., and Maguitman A.G., and Loui R.P., Logical models of argument, 337383, New York, NY, USA, ACM Comput. Surv., ACM Press, 4, 2000.

[CHR 09] Christian Borgelt,Matthias Steinbrecher, Rudolf Kruse, Graphical Models: Representations for Learning, Reasoning and Data Mining, seconde édition,2009.

[CLE 01] Cleret M., Le Beux P., Le Duff F., Les systèmes d'aide à la décision médicale, Les cahier du numérique, vol.2, N°2/2001, pp.125-154,2001.

[COO 87] Cooper G. F., Probabilistic inference using belief networks is np-hard. Knowledge Systems Laboratory, 1987.

[COO 90] Cooper, G. F., The computational complexity of probabilistic inference using bayesian belief networks. Artificial intelligence, 42(2) : pp.393–405, 1990.

[COR 03] Corset Franck. Aide à l'optimisation à la maintenance à partir de réseau bayésien et fiabilité dans un contexte doublement censuré, Thèse doctorat, Université Joseph-Fourier - Grenoble I, 2003.

[COR 02] Cornuéjols Antoine, Laurent Miclet, Apprentissage artificiel : Concepts et algorithmes, édition Eyrolles, pp 364-365,2002.

[DAR 03] Darmoni S. J., Titres et travaux informatique de santé, Sciences et Technologies de l'information et de la communication, 2003.

[DAR 13] Darwiche A., A differential approach to inference in Bayesian networks, Journal of the ACM (JACM), 50.10.1145/765568.765570, 2013.

[DEC 90] Dechter R., of Publication: Encyclopedia of Cognitive Science, Intelligence, 49(6195), 1990.

[DEG 91] Degoulet P., Fieschi M., Traitement de l'information médicale, Méthodes et applications hospitalières, Masson, 1991.

[DEM 77] Dempster A., Laird N., Rubin D., Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, B 39: 1-38, 1977.

[DJE 06] Djebbar A., Merouani H. F., Vers une modélisation de la base de cas par un réseau bayésien : Application d'aide au diagnostic des pathologies hépatiques,

6ème Conférence Francophone de MOdélisation et SIMulation- MOSIM'06, Maroc, 3-5 Avril, 2006.

[DJE 10] Djebbar A., Refai A., Merouani H.F., RB_Maint : Un modèle probabiliste pour la maintenance d'un système RàPC. Colloque sur l'Optimisation et les systèmes d'Information, COSI 2010- Algérie, Ouargla 18-20 Avril 2010.

[DRU 09] Druzdel, Marek J., Rapid modeling and analysis with QGeNIe. In Proceedings of the International Multiconference on Computer Science and Information Technology (IMCSIT-2009), pages 101-108, Mragowo, Poland, October 12-14, 2009.

[DRU 99] Druzdzel M. J., GeNIe: A development environment for graphical decision-analytic models, In Proceedings of the AMIA Symposium (p. 1206), American Medical Informatics Association, 1999.

[DRU 99] Druzdzel Marek J., SMILE: Structural modeling, inference, and learning engine and GeNIe: A development environment for graphical decision-theoretic models (intelligent systems demonstration), In Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99), pp 902–903, Menlo Park, CA, 1999. AAAI Press/The MIT Press.

[FAT 07] Fatiha NAOUAR, Algorithme approximatif pour le diagnostic médical basé sur un réseau bayésien possibiliste guidé par ontologie floue, thèse de master, Université de Sousse, 2007.

[FIN 96] Finn V.J., An introduction to Bayesian Networks, Eyrolles, Première Edition, 1996.

[FOM 13] Fomin, Fedor,Villanger V., Yngve, Subexponential parameterized algorithm for minimum fill-in, SIAM Journal on Computing, vol. 42, no 6, pp. 2197-2216, 2013.

[FUN 13] Fung Robert M., Chang K. C., Weighing and Integrating Evidence for Stochastic Simulation in Bayesian Networks, In Proc. Conf. Uncertainty in Artificial Intelligence, (2013, April).

[FUN 13a] Fung Robert, Del Favero, Brendan, Backward Simulation in Bayesian Networks, 10.1016/B978-1-55860-332-5.50034-1,2013

[GAU 14] Gautier Baissas, mise en place d'une étude pilote, de faisabilité et d'évaluation du logiciel d'aide au diagnostic des anémies : web-anemia, thèse de doctorat, l'université mixte de médecine et de pharmacie de Rouen, 2014.

http://medecine-pharmacie.univ-rouen.fr/informatique-medicale technologies-del-information-et-de-la-communication-16308.kjsp?RH=138224035996

[GEI 90] Geiger D., Verma T., and Pearl J., Identifying independance in bayesian networks, Networks, 20 : pp.507–534, 1990.

[GOL 00] Goldstein M. Hoffman K. B. B., Coleman R. W., Musen M. A., Tu S. W., Advani A., Shankar R., O'Connor M., Implementing clinical practice guidelines while taking account of changing evidence: ATHENA DSS, an easily modifiable decision-support system for managing hypertension in primary care, Proceedings / AMIA ... Annual Symposium. AMIA Symposium, pp.300–304, 2000.

[GUO 04] Guo H., Boddhireddy P., & Hsu W., An ACO algorithm for the most probable explanation problem, AI: Advances in Artificial Intelligence, pp.251-282. 2004.

[HAL 04] Hallouli kallid, reconnaissance de caractères par méthodes markoviennes et réseaux bayésien, Thèse de doctorat à l'école national supérieure des télécommunications, spécialité : signal et image, pp.87-130,05mai 2004.

[HEC 92] Heckerman, D. E., Horvitz E. J., and Nathwani, B. N., Toward normative expert systems: Part i the pathfinder project, Methods of Information in Medicine, 31: pp. 90-105, 1992.

[HEN 92] Hénaut A., Corvol P., Degoulet P., Nouvelle méthode de traitement de l'information en médecine, 1992

[HOF 05] Hoffer E.P., Feldman M.J., Kim R.J., Famiglietti K.T., Barnett G.O., DXplain: Patterns of use of a mature expert system, AMIA Annu Symp Proc. (488): pp.321-5. 2005.

[HOM 12] Homam Mohammad Alsun, Indexation guidée par les connaissances en imagerie médicale, Thèse de Doctorat, l'Université européenne de Brettagne, 2012.

[HOP 18] Hopfield, John J. "Neural Networks and Physical Systems with Emergent Collective Computational Abilities." Feynman And Computation. CRC Press, 7-19, 2018.

[HOR 14]. Horny M., Bayesian networks, Technical Report N°.5, 2014.

[HSU 02] Hsu W. H., Guo H., Perry B. B., & Thornton J. A., Bayesian Network Tools in Java (BNJ) Project Page, Kansas State University Laboratory for Knowledge Discovery in Databases 2002.

[JEN 90] Jensen F. V., Lauritzen S. L., Olesen K. G., Bayesian updating in causal probabilistic networks by local computations, 1990.

[JEN 90] Jensen F. V., Olesen K. G., Andersen S. K., An Algebra of Bayesian Belief Universes for Knowledge-Based Systems. Networks: Special Issue on Influence Diagrams, Vol.20, No. 5, August 1990, pp.637-659, 1990.

[JEN 96] Jenson F.V., Introduction to Bayesian Networks, UCL Press, London, England, 1996.

[KAW 05] Kawamoto K., Houlihan C. A., Balas E. A. et Lobach D. F., Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success, BMJ, Vol 330 (7494), pp.765, Avril 2005.

[KIM 83] Kim J., Pearl J., A computational model for combinedcausal and diagnostic reasoningin inference systems, Dans Proceedings IJCAI-83, pp. 190–193, Karlsruhe, Germany, 1983.

[KIN 67] King L. S., « What is a diagnosis? », JAMA: The Journal of the American Medical Association, vol. 202, no. 8, pp. 714 -717, nov. 1967.

[KJA 90] Kjærulff U., Triangulation of graphs--algorithms giving small total state space, 1990.

[KON 08] Kong G., Xu D. L., Yang J. B., Clinical decision support systems: a review on knowledge representation and inference under uncertainties, International Journal of Computational Intelligence Systems, vol. 1, no. 2, pp. 159-167, 2008.

[KON 93] Kononenko I., Inductive and Bayesian Learning in Medical Diagnosis Applied Artificial Intelligence, 1993.

[KRU 03] Kruegel Christopher, Darren Mutz, William Robertson, et Fredrik Valeur, Bayesian event classification for intrusion detection, Dans 19th Annual Computer Security Applications Conference, LasVegas. IEEE Computer Society, December 2003.

[KUM 05] Kumar V., Abbas A.K., Fausto N., Robbins and Cotran Pathologic Basis of Disease, 7th Edn., WB Saunders Co., New York, ISBN-10: 0721601871. pp 15-52, 2005.

[LAB 03] Labatut V., Réseaux causaux à grande échelle : un nouveau formalisme pour la modalisation du traitement de l'information cérébrale, thèse doctorat de l'université d'Aix-Marseille, 2003.

[LAU 88] Lauritzen S. L., Spiegelhalter D. J., Local computations with probabilities on graphical structures and their application to expert systems, Journal of the Royal Statistical Society. Series B (Methodological), pp.157-224, 1988.

[LEB 94] Lebleux P., Burgun A., Mireille C., De la méconnaissance à l'expertise, Laboratoire d'informatique médicale, Springer-Verlag, Paris, France, 1994.

[LEN 80] Lenoire P., Bourel M., Rouger J M., Chales G, Système d'aide au diagnostic médical : Méthodes utilisées, Med inform,vol 5, n°4, pp.291-307, 1980.

[LEP 92] Lepage E., Fieschi M., Traineau R., Gouvernet J. et Chastang C., Système d'aide à la décision fondé sur un modèle de réseau bayesien application à la surveillance transfusionnelle. Informatique et Santé, Vol 5, pp. 76-87, 1992.

[LER 02] Leray P., francois O., réseaux bayésiens pour la classification méthodologie et illustration dans le cadre du diagnostic médical, INSA Rouen/PSI, FRE CNRS 2645.BP 08-Av. de l'université.76801, St Etienne du Rouvray Cedex.2002

[LER 04] Leray Philippe, Olivier François, Etude Comparative d'Algorithmes d'Apprentissage de Structure dans les Réseaux Bayésiens, Journal électronique d'intelligence artificielle (JEDAI), 5(39) : pp.1–19, 2004.

[LOB 07] Lobach D. F., Kawamoto K., Anstrom K. J., Russell M. L., Woods P., Smith D., Development, deployement and usability of a point-of-care decision support system for chronic disease management using the recently-approved HL7 decision support service standard, Studies in Health Technology and Informatics, vol.129, pp.861-865,2007.

[LUC 04] Lucas P. J. F., van der Gaag L. C., Abu-Hanna A., Bayesian networks in biomedicine and health-care. Artificial Intelligence in Medicine 30, p p. 201–214, 2004.

[MAC 96] Mackay D., Introduction to Monto Carlo Methods. In M Jordan, Editor, Erice Summer School, 1996.

[MAD 05] Madsen A. L., Jensen F., Kjærulff U. B., and Lang M. Hugin - the tool for Bayesian networks and influence diagrams. International Journal on Artificial Intelligence Tools, 14(3):507–543, 2005.

[MAL 16] Malek M., Un modèle hybride de mémoire pour le raisonnement à partir de cas. Thèse de doctorat, Université Joseph Fourier, 30 octobre 1996.

[MCC 88] McCulloch W. S., Pitts W., A logical calculus of the ideas immanent in nervous activity, 1527,0262010976, Cambridge, MA, USA, MIT Press, 1988.

[MIL 09] Miller R. A., Computer-assisted diagnostic decision support: history, challenges, and possible paths forward. Advances in Health Sciences Education, Vol 14, pp. 106, 2009.

[MOK 16] Mokeddemm Sid Ahmed, Fouille de données pour l'annalyse de traces patientes. Thèse de doctorat, Laboratoire d'informatique d'Oran, L'Université d'Oran Ahmed Benbella, MAI 2016.

[MOR 15] Moreno Mathieu, Développement des systèmes d'aide à la décision dans les cabinets de médecine générale en France, Thèse de Doctorat, l'Université de Lorraine, 2015.

[NAI 04] Naim P., Wuillemin P., Leray H., Pourret P.O., et Becker A., Réseaux Bayésiens. Groupe Eyrolles, ISBN 2-212- 11371, 2004.

[NAï 07] Naïm Patrick, Pierre-Henri Wuillemin, Philippe Leray, Olivier Pourret, et Anna Becker, Réseaux bayésiens, 3ième édition, 2007.

[NAO 07]. Naouar Fatiha., Modélisation d'une détection de diagnostic d'une maladie par les réseaux bayésiens possibiliste, Institut Supérieur des Sciences Appliquées et de Technologie de Sousse Université de Sousse Tunisie, Séminaire SIM'07, 2007.

[PEA 86] Pearl J., Fusion, propagation, and structuring in belief networks. Artificial intelligence, 29(3) : pp.241-288, 1986.

[PEA 88] Pearl J., Probabilistic Reasoning in Intelligent Systems. Networks of Plausible Inference, San Mateo, CA, Morgan Kaufmann Publishers, 1988.

[PEA 88] Pearl J., Probabilistic reasoning in intelligent systems: networks of plausible inference, Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, 1988.

[PEA 98] Pearl J., Bayesian networks, 149153, 0262511029, Cambridge, MA, USA, The handbook of brain theory and neural networks, MIT Press, 1998.

[PEI 13] Pei, Bin, Zhao, Suyun, Chen, Hong, et al. FARP: Mining fuzzy association rules from a probabilistic quantitative database. Information Sciences, vol. 237, pp. 242-260, 2013.

[PEW 01] Pewsner D., Bleuer J., Bucher P., Battaglie H.C., Sur la voie de l'intuition Théorème de Bayes et Diagnostic en médecine générale, pp. 41-45, 2001.

[QIN 05] Qin Xinzhou, A probabilistic-based framework for infosec alert correlation, PhD thesis, Atlanta, GA, USA, 2005.

[RAI 07] Riascos L.A.M, Simoes M.G., Miyagi P.E., A Bayesian network fault diagnostic system for proton exchange membrane fuel cells, Journal of Power Sources, Elsevier, 2007.

[REF 11] Refai Ahlem, Merouani Hayet Farida, Maintenance d'un réseau bayésien pour le diagnostic des pathologies des seins, la deuxième édition de la conférence internationale sur les systèmes et le traitement de l'information (icsip'11), l'université de Guelma 8 mai 1945, 15-17 mai 2011.

[REF 12] Refai Ahlem, Merouani Hayet Farida, Application à l'aide de diagnostic médicale à base de réseau bayésien dans le cadre de la maintenance, la troisième édition de la conférence internationale sur l'image et le traitement du signal et leurs applications à Mostaganem 2012.

[SAH 07] Sahi F., Yavuz M. C., Arnavut Z., Uluyol O., Fault diagnosis for airplane engines using Bayesian networks and distributed particle swarm optimization, Parallel Computing, Elsevier, 2007.

[SAH 98] Sahami Mehran, Susan Dumais, David Heckerman, et Eric Horvitz, A Bayesian Approach to Filtering Junk E-mail, Dans AAAI Workshop on Learning for Text Categorization, July 1998.

[SCH 82] Schank R., Dynamic Memory: A Theory of Learning in Computers and People (New York: Cambridge University Press, 1982), New York: Cambridge University Press, 1982.

[SCH 86] Schank R., Explanation Patterns: Understanding Mechanically and Creatively, Erlbaum, 1986.

[SER 01] Séroussi B., Bouaud J., and Antoine E.C., ONCODOC: a successful experiment of computer-supported guideline development and implementation

in the treatment of breast cancer. Artificial Intelligence in Medicine, 22(1): pp.43–64, 2001.

[SER 13] Seroussi B., Le Beux P., & Venot A., L'aide au diagnostic médical. In Springer Verlag France (Ed.), Informatique médicale, santé. Fondements et applications 1st ed, pp.147–175, Paris : Springer, 2013.

[SER 99] Séroussi B., Bouaud J., Antoine, E. C., Users' evaluation of OncoDoc, a breast cancer therapeutic guideline delivered at the point of care, Proceedings / AMIA ... Annual Symposium. AMIA Symposium, pp. 384–389, 1999.

[SHA 13] Shachter, Ross D., Peot Mark Alan, Simulation approaches to general probabilistic inference on belief networks, arXiv preprint arXiv:1304.1526, 2013.

[SHA 90] Shachter R. D, D'Ambrosio B, Del Favero B, Symbolic Probabilistic Inference in Belief Networks, In AAAI, Vol. 90, pp. 126-131, 1990.

[SID 13] Sidoine Pierre V., DONFACK GUEFACK, Modélisation des signes dans les ontologies biomédicales pour l'aide au diagnostic, Thèse / Université de Rennes 1, 2013.

[SOU 95] Sournia J.-C, Histoire du diagnostic en médecine. Santé, 1995.

[SPI 98] Spiegelhalter D. J: Bayesian Graphical Modelling: A Case-Study in Monitoring Health Outcomes. Applied Statistics 47, Part 1, p. 115-133, 1998.

[THI 99] Thijssen C. R. T., SmileX: An ActiveX decision-analytic reasoning engine and its application to evaluation of credit applicants, Master's thesis, Delft University of Technology, 1999.

[WAN 10] Wang, Chengdong et MOSLEH, Ali. Qualitative-Quantitative Bayesian Belief Networks for reliability and risk assessment. In:Proceedings-Annual Reliability and Maintainability Symposium (RAMS). IEEE, pp. 1-5,2010.

[WIL 07] Wilkinson Darren J., Bayesian methods in bioinformatics and computational systems biology, Journal of Briefings in bioinformatics, 8(2): pp. 109–116, April 2007.

[YUA 17] Yuan, Xiuli. An improved Apriori algorithm for mining association rules. In: AIP conference proceedings. AIP Publishing, pp. 080005, 2017.

[ZHA 94] Zhang N. L., Poole D., A simple approach to Bayesian network computations, In Proc. of the Tenth Canadian Conference on Artificial Intelligence, 1994.

[ZHA 94] Zhaoyu Li, Bruce d'Ambrosio, Efficient inference in Bayes networks as a combinatorial optimization problem, International Journal of Approximate Reasoning 11.1: pp. 55-81, 1994.

[ZHO 05] Zhou Lina, Jinjuan Feng, Andrew Sears, Yongmei Shi, Applying the Naïve Bayes Classifier to Assist Users in Detecting Speech Recognition Errors, Dans Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05), Washington, DC, USA, janury 2005.

[ZOG 11] Zoghlami Aymen, Approche probabiliste pour l'analyse de l'impact des changements dans les programmes orientés objet, Mémoire (M.Sc.), 2011.

1. ***Mamelon*** n.m saillie centrale du sein, ou débouchent les canaux galactophores, et qui permet au nourrisson de téter (V. ALLAITEMENT). [Le mamelon est entouré par l'aréole, surface annulaire pigmentée d'étendue variable.]
2. ***Opacité*** n.m. petit lobe. Subdivision d'un lobe. Segment arrondi et systématisé d'un organe désigne une unité structurale et fonctionnelle d'un organe en générale arrondi, étant séparé de son voisin par peu de tissu conjonctif qui en forme la limite...
3. ***Masse*** quantité relativement grande de substance solide ou pâteuse, qui n'a pas de forme.
4. ***Adénofibrome*** c'est une tumeur bénigne du sein qui se développe aux dépens du tissu conjonctif on l'appelle aussi fibroadénome car elle est de nature fibreuse.
5. ***Kyste*** est une cavité anormale remplie de liquide ou de semi-liquide qui se forme dans un organe ou un tissu. La grande majorité des kystes sont bénins, c'est-à-dire non cancéreux. Toutefois, ils peuvent perturber le fonctionnement d'un organe ou causer des douleurs.
6. ***Lipome*** est une **tumeur bénigne constituée de graisse** qui n'entraîne généralement aucune complication. Il est avant tout gênant pour la personne atteinte sur le plan esthétique lorsqu'il est situé au niveau de la peau.
7. ***Mastodynie*** Les douleurs au sein, ou Mastodynie, sont en général d'origine hormonale**.**
8. ***Adénopathie*** Il s'agit d'une atteinte des ganglions lymphatiques, sans qu'il soit possible de préciser immédiatement l'origine de l'affection qui les modifie. Dans le langage courant, le synonyme immédiat d'adénopathie est "ganglion"

Graphe orienté sans circuits

Dans cette section nous rappelons quelques concepts élémentaires de la théorie des graphes nécessaires pour la compréhension de la thèse. L'utilisation des graphes et de leurs algorithmes s'est généralisée pour devenir un des outils mathématiques de base pour la modélisation et la résolution de très nombreux problèmes scientifiques et techniques.

En effet les graphes sont un outil essentiel pour représenter les modèles probabilistes ou autres utilisations de l'intelligence artificielle et des systèmes experts. Beaucoup de résultats théoriques et algorithmiques de la théorie des graphes peuvent être utilisés pour analyser les différents aspects de ces domaines.

En revanche, nous avons maintenu le vocabulaire usuel en théorie des réseaux bayésiens, qui fait parfois double emploi avec celui de la théorie des graphes et qui fait largement usage des analogies avec la généalogie : par exemple nous parlons de « parent » et non de « prédécesseur » ainsi que « d'enfant » et non de « successeur » ; de même nous désignons par le terme « racine » un nœud sans prédécesseurs, et par le terme « feuille » un nœud sans successeurs, même si le graphe oriente considère n'est pas une arborescence.

A.1 Préliminaire sur les graphes orientés sans circuits

A.1.1 Graphe sans circuits

Un graphe orienté est un couple (I, G) ou I est un ensemble fini, et G est une partie de l'ensemble des couples (i, j) d'éléments de I.

Si cela ne prête pas à confusion, nous pourrons dire « le graphe G » (en précisant éventuellement « sur I ») pour designer « le graphe (I, G) ».

Les éléments de I sont appelés nouds, ou sommets.

Les éléments de G sont appelés arcs.

Un chemin, relativement à G, est une suite d'éléments, de I, (i0, . . ., ik) (ou $k \geq 1$) vérifiant :

$\forall s \in \{1, \ldots, k\} \{ i_{s-1}, i_s \} \in G$; i_0 est l'origine du chemin et i_k son extrémité.

Un circuit est un chemin dont l'origine et l'extrémité sont identiques. Les graphes orientés sans circuits, en particulier sans boucles pour tout i : $i,(i,i) \notin G$

Voici un exemple sur lequel toutes les notions et notations dans la suite de cette annexe :

:

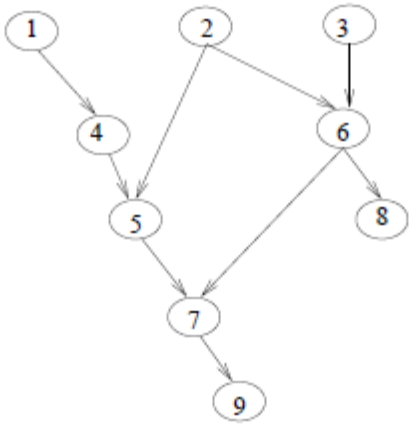


FIG. A.1 – Exemple d'un graphe orienté sans circuits (I, G).

I = {1, 2, 3, 4, 5, 6, 7, 8, 9}.
G = {(1, 4), (2, 5), (2, 6), (3, 6), (4, 5), (5, 7), (6, 7), (6, 8), (7, 9)}.

Remarque : On distinguera bien le vocabulaire des graphes orientés sans circuits de celui des graphes non-orientés : un graphe non-orienté est un couple (I, H) ou H est un sous ensemble de l'ensemble des singletons {i} et des paires {i, j} d'éléments de I distincts (une paire est une partie à deux éléments).
A tout graphe orienté sans circuits (I, G) on peut associer le graphe non orienté sous-jacent (I, H) ou H est l'ensemble des paires {i, j} telles que l'un des couples (i, j) ou (j, i) appartienne à G (un seul de ces deux couples peut appartenir `à G en raison de l'absence de circuits).
Les éléments de H sont appelés arêtes.
Dans un graphe non-orienté on appelle chaîne toute suite $(i_0, \ldots, i_k)$ (ou $k > 1$) vérifiant

$\forall s \in \{1, \ldots, k\} \{ i_{s-1}, \ldots, i_s \} \in H$ et cycle une chaîne dont l'origine et l'extrémité sont identiques.
Il peut se produire que dans le graphe non-orienté sous-jacent au graphe orienté sans circuits existent des cycles (dans l'exemple ci-dessus, on a le cycle (2, 5, 7, 6, 2)).
On doit signaler ici un risque de confusion entre le vocabulaire français et le vocabulaire

Anglais, ou circuit se dit « cycle » et un graphe orienté sans circuits se dit « Directed Acyclic Graph » (DAG).

A.1.2 Parties caractéristiques associées à un nœud.

Etant donné un graphe orienté sans circuits (I, G) on associe à tout élément i de I :

1. l'ensemble p(i) de ses parents (aussi appelés prédécesseurs ou antécédents immédiats, ensemble noté également, en théorie des graphes, $\Gamma^{-}(i)$ ou $N^{-}(i)$)

$$p(i)=\{j\ ;(i,j)\in G\}$$

2. l'ensemble e(i) de ses enfants (aussi appelés successeurs, ou successeurs immédiats, ensemble noté également, en théorie des graphes, $\Gamma^{+}(i)$ ou $N^{+}(i)$)

$$e(i)=\{j\ ;(i,j)\in G\}$$

3. l'ensemble ℓ(i) dit lignée ascendante de i, aussi appelé ensemble de tous ses antécédents, qui s'obtient par itération de la relation "être parent" :

$$l(i)=\{j\ ;\exists(l_0,\dots,l_k)\ l_0=i\ l_k=j,\forall\ s\in\{1,\dots,k\}\ (ls_{-1},\ l_s)\}\in G$$

4. l'ensemble d(i) de ses descendants, aussi dit descendance, ou lignée descendante, de i qui s'obtient par itération de la relation « être enfant » :

$$d(i)=\{j\ ;\exists(l_0,\dots,l_k)\ l_0=i\ l_k=j,\forall\ s\in\{1,\dots,k\}\ (ls_{-1},\ l_s)\}$$

5. l'ensemble c(i) des collatéraux à i, qui est l'ensemble des éléments, autres que i lui-même, qui ne sont ni dans sa lignée ni dans sa descendance :

$$c(i)=I-(\{i\}\cup l(i)\cup d(i))$$

6. l'ensemble a(i) des aïeux ou ancêtres de i, qui est l'ensemble des éléments de la lignée ascendante de i autre que ses parents :

$$a(i)=l(i)-p(i)$$

Une racine d'un graphe est un nœud sans parent (les sommets 1, 2 et 3 du graphe de l'exemple sont ses racines) alors qu'une feuille est un nœud sans enfant (les nouds 8 et 9 du graphe de l'exemple sont ses feuilles).

En théorie des graphes, les termes « racine » et « feuille » sont plutôt réserves au cas des arborescences ; dans le cas des graphes orientés les plus généraux, les termes correspondants seraient alors « source » et « puits ».

Remarque : Certains de ces ensembles peuvent bien sûr être vides.

Exemple :

Dans l'exemple on a pour l'élément 5 :

1. p (5) = {2, 4}.

2. ℓ (5) = {1, 2, 4}.

3. a (5) = {1}.

4. e (5) = {7}.

5. d (5) = {7, 9}.

6. c (5) = {3, 6, 8}.

A.2 Numérotages des nœuds.

Soit I un ensemble fini tel que Card (I) = n.

Un numérotage des nœuds de I est une suite (i_1, . . ., i_n) telle que {i_1, . . ., i_n} = I

A.2.1 Numérotation hiérarchique (aussi appelé ordre hiérarchique)

Définition

Une numérotation (i1, . . ., in) est dite hiérarchique pour G si :

$$s \in \{1,\ldots, n\}\ p\ (i_1) \subset \{i1,\ldots, i_{s-1}\}$$

Autrement dit tout élément est classé après ses parents.

Remarque :

Il résulte immédiatement de la définition que :

1. $s \in \{1,\ldots, n\}\ \ell\ (i_1) \subset \{i1,\ldots, i_{s-1}\}$

   C'est à dire que tout élément est classé après tous ceux de sa lignée.

1. $\forall\ s \in \{1,\ldots, n\}\ d\ (i_1) \cap \{i1,\ldots, i_{s-1}\} = \varnothing$,
2. En particulier : p (i1) =∅

4. L'absence de circuits (en anglais : acyclicity) est une condition suffisante d'existence de numérotation hiérarchique, autrement dit, si dans un graphe il n'existe pas de numérotation hiérarchique, il existe des circuits ; c'est le cas, par exemple, du graphe suivant :

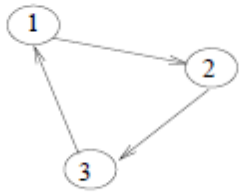


FIG. A.2 – Exemple d'un graphe avec circuit.

Production scientifique

Communications internationales

- Djebbar Akila, Refai Ahlem, Merouani Hayet Farida, RB_Maint : Un modèle probabiliste pour la maintenance d'un système RàPC. Colloque sur l'Optimisation et les systèmes d'Information, COSI 2010- Algérie, Ouargla 18-20 Avril 2010.
- Refai Ahlem, Merouani Hayet Farida, Maintenance d'un réseau bayésien pour le diagnostic des pathologies des seins, la deuxième édition de la conférence internationale sur les systèmes et le traitement de l'information (icsip'11), l'université de Guelma 8 mai 1945, 15-17 mai 2011.
- Refai Ahlem, Merouani Hayet Farida, Application à l'aide de diagnostic médicale à base de réseau bayésien dans le cadre de la maintenance, la troisième édition de la conférence internationale sur l'image et le traitement du signal et leurs applications à Mostaganem 2012.

Communication nationale

- Refai Ahlem, Merouani Hayet Farida, Maintenance d'un réseau bayésien : application au diagnostic médical, Doctoral consortiumin computer science, JDI, 2017.

Publications internationales

- Refai Ahlem, Merouani Hayet Farida, Aoures Hayet, Maintenance of bayesian network: application using a medical diagnosis, of international journal Evolving Systems,Vol 7, N°3, pp 187–196, 2016.