Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

**BADJI MOKHTAR -ANNABA**
**UNIVERSITY**
**UNIVERSITE BADJI MOKHTAR**
**ANNABA**

**جامعة باجي مختار**
**- عنابة -**

**Faculté des Sciences de l'Ingénieur** **Année : 2013**
**Département d'Informatique**

**THESE**

Présentée en vue de l'obtention du diplôme de
**DOCTORAT** en Informatique

# Optimisation de la recherche d'un cas Bayésien

**Option :**
Informatique

Par
**Akila DJEBBAR-ZAIDI**

**Directeur de thèse :**

| | |
|---|---|
| MEROUANI Hayet Farida | MCA. Univ. Badji Mokhtar-Annaba |

**Devant le jury**

**Président** :

| | |
|---|---|
| SERIDI Hamid | Prof. Univ. 8 Mai 1945-Guelma |

**Examinateurs** :

| | |
|---|---|
| MOUSSAOUI Abdelouahab | Prof. Univ. Ferhat Abbes-Sétif |
| KHOLLADI Mohamed-Khireddine | Prof. Univ. Mentouri-Constantine |
| BENBLIDIA Nadjia | MCA. Univ. Saad Dahlab-Blida |
| TLILI Yamina | MCA. Univ. Badji Mokhtar-Annaba |

***Remerciements***

*Je remercie vivement toutes les personnes qui m'ont aidé, encouragé tout au long de ce travail de thèse. Je tiens à remercier en particulier les membres du jury:*

*Madame Hayet Farida MEROUANI, maître de conférences à l'Université Badji Mokhtar Annaba, pour m'avoir accueilli dans son groupe de recherche, pour avoir encadré mon travail, pour sa confiance et sa patience. Enfin je la remercie pour m'avoir permis de réaliser cette thèse dans les meilleures conditions possibles.*

*Monsieur Hamid Séridi, professeur à l'Université 8 Mai 1945 à Guelma, pour avoir accepté de présider ce jury et d'avoir voulu juger ce travail.*

*Monsieur Abdelouahab Moussaoui, professeur à l'Université Ferhat Abbes à Sétif, pour avoir accepté d'examiner mon travail et de participer au jury.*

*Monsieur Mohamed-Khireddine Kholladi, professeur à l'université Mentouri à Constantine, qui a accepté d'évaluer ce travail et de participer au jury de soutenance.*

*Madame Yamina Tlili, maître de conférences à l'Université Badji Mokhtar Annaba, pour avoir accepté d'être rapporteurs de ce travail et pour leurs remarques qui ont contribué à l'amélioration de cette thèse.*

*Madame Nadjia Benblidia, maître de conférences à l'Université Saad Dahlab-Blida, d'avoir répondu à notre demande pour faire partie de ce jury en tant que examinatrice de ce travail.*

*Monsieur Ghizil Mohamed, medecin radiologue a l'hôpital iben rochd d'Annaba qui a dirige ce travail et qui a permis, grace a ses competences, de le mener a terme.*

*Monsieur Farouk Meddeh et melle Fatima Zohra Amiri pour leurs aides, leurs gentillesses, et pour l'interet qu'ils ont porte a ce travail.*

*Enfin, je remercie toute personne ayant collaboré de pré ou de loin à la réalisation de ce travail en particulier mon conjoint qui m'a soutenu et encouragé.*

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

***A la mémoire de mes parents,***

***A mon mari,***
***à mes enfants: Salsabil et Anfel,***
***à toute la famille.***

# ملخـص

التفكير بالحالات عبارة عن طريقة لحل مشاكل المؤسسة على إعادة الاستعمال بالتماثل التجارب السابقة تدعى "حالة". البحث عن الحالات هي مرحلة أولى في محور التفكير بالحالات و قياس التشابه يلعب دورا مهما في البحث عن الحالات .مرحلة التكييف تعتبر أصعب المراحل في محور التفكير بالحالات. الفكرة الأساسية تشمل تشكيل قاعدة الحالات بواسطة شبكة بايز . إن الشبكات البايزية من أحسن وسائل الترددية عن طريق العرض البياني الواضح و قوانين الاحتمالات الشرطية المعرفة على هذا البيان . نهتم بتشكلية مرحلة التذكر و مرحلة التكييف . في الواقع ، مرحلة التذكر تسعى لاختيار الحالات المشابه للنموذج اللوخطي الذي يعتبر الشبكة البايزية مثل نموذج اللوخطي . مرحلة التكييف تسعي لتوليد حل لمسألة المصدر بواسطة المسألة المتذكرة . العلاقة بين هاتين الحالتين معرفة بواسطة مقياسين : مقياس التشابه و مقياس التكييف ، هذه العلاقة تهدف لضمان الحصول على الحالة الأكثر قابلية للتكييف من أجل تحسين نتائج التفكير بالحالات . نتائج استعمال الشبكات البايزية قيمت بواسطة قاعدة الحالات للأمراض الكبيدية الأكثر انتشار . النتائج المحصل عليها مرضية و تبرهن على فعالية الشبكات البايزية مثل نطاق لتمثيل المعلومات الطبية و مثل وسيلة لتفكير التشخيصي .

**المفاتيح :**

التفكير بالحالات ، التذكر، التكييف ، قياس التشابه ، الشبكة البايزية ، التشخيص المرضي.

# Résumé

Le Raisonnement à Partir de Cas (RàPC) est une approche de résolution de problèmes basée sur la réutilisation par analogie des expériences passées nommées « cas ». La remémoration des cas est une étape primaire dans le cycle RàPC et la mesure de similarité joue un rôle très important dans la recherche des cas. L'étape d'adaptation est considéré une des étapes les plus difficiles du cycle RàPC. Nous sommes intéressés à la modélisation de la phase de la remémoration et la phase d'adaptation appliquée au diagnostic des pathologies hépatiques. L'idée principale consiste à modéliser la base de cas par un réseau Bayésien. Les réseaux Bayésiens sont d'excellents outils de modélisation de l'incertain grâce à leur représentation graphique claire et aux lois de probabilités conditionnelles définies sur ce graphe. En effet, la phase de remémoration consiste à sélectionner le cas le plus similaire à travers d'une mesure de similarité Log-linéaire. Par ailleurs, la phase d'adaptation consiste à fournir une solution cible au problème cible à partir du cas remémoré. La dépendance entre ces deux phases est définie par deux mesures à savoir: une mesure de similarité et une mesure d'adaptation. Cette dépendance a pour objectif de garantir la recherche d'un cas qui soit le plus facile à adapter afin d'améliorer la performance du RàPC. Les performances de l'utilisation des Réseaux Bayésiens sont évaluées sur une base de cas des maladies du foie les plus fréquentes. Les résultats obtenus sont satisfaisants et montrent l'efficacité de réseaux Bayésiens comme un cadre de représentation des connaissances médicales et comme outil de raisonnement diagnostique.

**Mots clés**

Raisonnement à Partir de Cas (RàPC), remémoration, adaptation, mesure de similarité, réseau Bayésien (RB), diagnostic médical.

# Abstract

Case Based Reasoning (CBR) is an approach of solving problem which is based on the reuse, by analogy, of past experiences called case. Retrieval of cases is a primary step in CBR, and the similarity measure plays a very important role in case retrieval. The Adaptation step is considered to be one of the most difficult parts of the CBR cycle. We are interested to the retrieval and adaptation phases for a CBR applied to the diagnosis of hepatic pathologies. The main idea consists in a modelling the case base by a Bayesian Network (BN). Its are excellent tools for modelling the uncertainty in terms of their clear graphic representation as well as the conditional probabilities laws defined on a graph. The retrieval phase consists in selecting the most similar case of log linear model witch consists of considering similarity measure as a log-linear model. The adaptation phase means modifying solutions of retrieved cases to fit the current problem. The dependence between these two phases is definite by two measures: a similarity measure and an adaptation measure. This dependence has for objective to guarantee the retrieved case which is the easiest to adapt and improve the performance of CBR. The performances of the use of the Bayesian Networks are estimated by a case base of the most frequent diseases of the liver. The obtained results are satisfactory and show the efficiency of Bayesian networks as a frame of representation of the medical knowledge and as tool of diagnostic reasoning.

**Keywords**

Case Based Reasoning (CBR), retrieval, adaptation, similarity measure, Bayesian Network (BN), medical diagnosis.

# Table des matières

# Liste des Figures

# Liste des Tables

# Travaux publiés dans le cadre de cette thèse

**Revues internationales**

Djebbar A and Merouani H.F., Optimizing retrieval phase in CBR through Pearl and JLO algorithms for medical diagnosis. International Journal of Advanced Intelligence Paradigms IJAIP, Inderscience, Vol. 4, N° 3/4, 2012. ISSN: 1755-0386, 2012.

Djebbar A and Merouani H.F., Retrieval and adaptation in CBR through Bayesian Network for diagnosis of hepatic pathologies. International Journal of Hybrid Intelligent System IJHIS, IOS press, Vol. 9, No.3, 01 November 2012. ISSN: 1448-5869, pp. 123-134.

Djebbar A and Merouani H.F., Applying BN in CBR Adaptation-Guided Retrieval for medical diagnosis. International Journal of Hybrid Information Technology, IJHIT, Vol. 5, No.2, April, 2012. ISSN: 1738-9968, pp. 41-56.

Djebbar A and Merouani H.F., Retrieval of cases by using a Bayesian Network applied to the medical diagnosis. The Mediterranean Journal of Computers and Networks Medjcn. 2012. Vol 8 N° 2, April 2012, ISSN: 1744-2397, pp. 69-74.

**Conférences internationales**

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

Djebbar A and Merouani H.F. Une modélisation bayésienne de la remémoration pour l'aide au diagnostic. International conference on Information Technology and e-Service ICITeS 2011, pp 254-257, ISBN: 978-9938-9511-0-3, April 10-12, Sousse, Tunisia.

Djebbar A and Merouani H.F**.** BN-CBR model for diagnosis of hepatic pathologies, Biomedical Engineering International Conference (BIOMEIC 2012), October 10-11, Tlemcen, Algeria. ISSN: 2253-0886.

**Conférences nationales**

Djebbar A and Merouani H.F. Un modèle statistique de la remémoration d'un RàPC pour l'aide au diagnostic médical. 1$^{ères}$ Journées Doctorales du Laboratoire d'Informatique d'Oran JDLIO 2011, 31 Mai et 01 Juin 2011, Oran, Algérie.

Djebbar A and Merouani H.F. L'intitulé du papier est : Une Remémoration Bayésienne d'un RàPC pour l'Aide au Diagnostic Médical. Rencontres sur la Recherche en Informatique R$^2$I 2011, 12-14 Juin 2011, Tizi-Ouzou, Algérie.

# Introduction Générale

## 1. Contexte et problématique

La construction d'un système capable de reproduire les activités de raisonnement de l'être humain représente le rêve des chercheurs travaillant en intelligence artificielle. C'est la raison pour laquelle la conception des systèmes à partir de cas capables de réaliser des fonctions de raisonnement symbolique constitue actuellement un champ primordial des recherches. De tels systèmes 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 domaine de notre travail est la médecine où la recherche fondamentale en psychologie cognitive a été longtemps basée sur l'étude de situations de résolution de problèmes bien délimitées et relativement fermées, c'est-à-dire des situations dans lesquelles le problème est clairement posé au départ et le sujet doit chercher une stratégie pour aboutir à un but donné. L'analyse du raisonnement diagnostique des médecins, comme l'étude de l'expertise d'autre professionnels (joueurs d'échec, pilotes de ligne, etc.), a fourni l'occasion d'élaborer et de tester des hypothèses concernant des formes de Résolution de Problème *(probleme solving)* plus complexe, impliquant des démarches d'inférence et d'investigation à l'intérieur de systèmes ouverts et multidimensionnels, dans lesquels la définition même du problème pose souvent des difficultés.

La résolution de problème dans le domaine médical montre que la prise en compte de symptômes cliniques, dépend de la structure des hypothèses diagnostiques formulées par le médecin, dés les premières minutes de son contact avec le patient. L'analyse des facteurs qui provoque certains biais ou erreurs du raisonnement diagnostique peut faciliter la création et l'application de procédures de contrôle plus adéquates dans la gestion des services médicaux.

Parmi les différents modes de raisonnement, capables de transcrire le raisonnement humain sous des formes exploitables par ordinateur, un mode est particulièrement adapté aux systèmes d'aide au diagnostic: le Raisonnement à Partir de Cas (RàPC[1]).

Le raisonnement à partir de cas est une approche récente de résolution de problèmes qui est fondamentalement différentes de la plupart des approches utilisées en IA[2] [AAM 94]. Le RàPC est capable d'utiliser une connaissance bien spécifique retenue depuis des expériences précédentes sur des situations bien précises (cas). Un nouveau problème est résolu en cherchant des cas similaires dans le passé, et sa réutilisation dans la nouvelle situation. Une deuxième grande différence est que le RàPC est une approche qui incrémente et accepte l'apprentissage. En effet, les nouvelles expériences sont retenues, une fois le problème résolu, ainsi on enrichi les connaissances et on rend cette expérience valable pour la résolution de nouveaux problèmes.

Afin d'améliorer les performances du système RàPC, le raisonnement utilise des cas déjà stockés dans une base de cas. Celle-ci est supposée représentative de l'ensemble des problèmes susceptibles d'être posés au système. Mais plus la base s'accroît, et plus le temps de calcul sera long. C'est pourquoi les techniques d'organisations de la mémoire et les algorithmes de la recherche et d'appariement sont particulièrement important, d'où plusieurs organisations sont présentées dans [KOL 93].

Dans le cadre de notre étude, nous nous intéressons à la phase de la remémoration ainsi qu'à la phase d'adaptation. Une modélisation de la remémoration en raisonnement à partir de cas qui repose sur un réseau bayésien (BN[3]) via un modèle log linéaire pour la représentation de la mesure de similarité est aussi présentée. La phase de remémoration consiste à sélectionner le cas le plus similaire du modèle Log-linéaire afin d'améliorer cette phase et rendre ainsi la phase d'adaptation plus facile. Nous proposons un algorithme d'adaptation avec une mesure d'adaptation qui s'appui sur les paramètres du réseau bayésien représentant la mémoire de cas.

[1] RàPC : *Raisonnement à Partie de Cas,* en anglais : *Case Based Reasoning*
[2] IA : *Intelligence Artificielle*
[3] BN : *Bayesian Network*

## 2. Motivations et objectifs

Dans le domaine médical, le raisonnement désigne les stratégies utilisées par les médecins dans l'objectif d'établir un diagnostic en s'appuyant sur les données hétérogènes disponibles extraites des systèmes d'acquisition. A titre d'exemple, les données médicales peuvent être les résultats d'un examen clinique, une image, un résultat de laboratoire, un signal, une séquence vidéo, etc. Lorsqu'on traite des données du monde réel, comme les données médicales, nous ne pouvons pas occulter l'aspect lié à l'imperfection affectant ces données. En effet, les données médicales souffrent, en général, au moins d'un type d'imperfection comme par exemple l'imprécision, l'incertitude, ou encore, les données manquantes. Pour ces raisons, l'aspect incertitude doit être pris en considération dans l'élaboration des systèmes destinés à apporter une aide au diagnostic qui sera réalisé par les médecins [PEW 01] [ALS 12].

Les modèles probabilistes, tels que les réseaux bayésiens sont largement utilisés dans les domaines d'aide à la décision. Ce sont des modèles représentant des connaissances incertaines du domaine traité et s'adaptant bien à la variabilité des observations.

L'approche proposée se base sur la modélisation de la base de cas par un réseau bayésien. Ainsi que sur la modélisation de la phase de la remémoration et de l'adaptation du système RàPC. Pour le domaine qui nous préoccupe, à savoir l'aide au diagnostic des pathologies hépatiques, l'utilisation des réseaux bayésiens s'avère être très bénéfique comme on va le voir dans le quatrième et le cinquième chapitre.

Cette modélisation repose sur la description, par des graphes, des relations de causalité existantes entre les variables définissant le domaine des maladies du foie. A chaque variable, on associe une distribution de probabilité. Pour calculer les probabilités conditionnelles, nous avons utilisés l'algorithme Pearl afin de bien mener l'inférence. Ces probabilités sont utilisées comme mesure de similarité, en sélectionnant le cas le plus probable. De plus, cette formalisation de la mémoire de cas offre la possibilité de modéliser la phase de remémoration et la phase d'adaptation afin d'obtenir un temps de calcul réduit et qui produit un diagnostic le plus précis possible et avec la plus grande certitude.

Ce présent travail doit être conçu en tenant compte de :

- La modélisation des connaissances
  Comment le système va présenter et formaliser les connaissances du domaine ?
- La remémoration des connaissances
  Comment le système va sélectionner et chercher les connaissances du domaine ?
- La réutilisation de la connaissance rencontrée.
  Comment le système va sélectionner, évaluer et adapter la connaissance rencontrée au contexte afin de pouvoir la réutiliser ?
- L'exploitation des connaissances du système.
  Comment le système pourra-t-il prendre en compte de nouvelles situations (connaissance) ?

Les objectifs de notre travail se résument aux points suivants :

- Construire une base de cas à l'aide d'un réseau bayésien afin d'obtenir un diagnostic plus efficace et de bonne qualité.
- Améliorer les performances du système RàPC afin d'obtenir:
  1. une remémoration plus efficace et plus précise : pour cela nous avons utilisé deux algorithmes :
     - L'algorithme Pearl
     - L'algorithme Extented-Pearl proposé.
  2. une adaptation facile à partir de la phase de remémoration.
- Etablir une étude comparative entre le modèle proposé et les autres modèles existants.

## 3. Organisation de la thèse

Outre l'introduction générale et la conclusion générale, le manuscrit est structuré en cinq chapitres. Pour situer le contexte de notre travail et pour sa bonne compréhension, trois chapitres seront dédiés à l'état de l'art. Les deux derniers chapitres sont consacrés à la présentation de notre contribution.

Le premier chapitre introduit les principes fondamentaux du RàPC et la notion de diagnostic médical. Nous donnons quelques définitions, des concepts de l'approche RàPC et des techniques d'organisations de la base de cas. Nous présentons un état de l'art des systèmes médicaux en RàPC. Une discussion est établie sur les différences qui existent entre les différents modèles de mémoire de cas.

Le deuxième chapitre est consacré à la présentation détaillée des réseaux bayésiens, ainsi que leurs propriétés notamment la représentation de l'indépendance conditionnelle. Nous explicitons les algorithmes d'inférence d'une manière générale et l'inférence exacte en particulier. Il s'agit, en fait, de l'algorithme de propagation des messages locaux utilisé dans les polyarbres et l'algorithme de l'arbre de jonction utilisé dans tout type d'arbres.

Le troisième chapitre est dédié à la présentation des différentes architectures sur l'intégration des réseaux bayésiens dans le Raisonnement à Partir de Cas et un état de l'art des systèmes BN/CBR. Nos choix et nos démarches adoptés pour la mise en place de notre système d'aide au diagnostic médical par le RàPC sont aussi présentés.

Le quatrième chapitre présente notre approche qui intègre le BN dans le RàPC pour l'aide au diagnostic des pathologies hépatiques. Nous nous intéressons particulièrement aux phases de remémoration et d'adaptation. En effet, nous proposons deux algorithmes, le premier dédié à optimiser la phase de remémoration et le deuxième pour faciliter la phase d'adaptation. Pour cela nous créons un lien entre ces deux phases grâce à deux mesures utilisées à savoir : une mesure de similarité log linéaire et une mesure d'adaptation. Cette dernière permet de sélectionner le cas le plus facilement adaptable.

Le cinquième chapitre fournit les résultats expérimentaux des deux algorithmes développés dans le quatrième chapitre. Nous montrons comment ces algorithmes améliorent les performances du RàPC et nous établissons une comparaison entre l'approche proposée et d'autres approches.

Enfin, nous terminons cette thèse par une conclusion dans laquelle, nous dressons un bilan de notre travail et nous proposons quelques perspectives possibles pour poursuivre cette recherche.

# Chapitre 1

# Le Raisonnement à Partir de Cas et le diagnostic médical

## Sommaire

# 1. Introduction

Le raisonnement en Intelligence Artificielle a souvent été synonyme d'inférence par règles ou par modèles. Et ce malgré l'existence d'une grande diversité de modes de raisonnement à l'image de la richesse des mécanismes mentaux de l'homme. La majeure partie des applications en Intelligence Artificielle consiste à reproduire le raisonnement humain. Le raisonnement par cas est une approche de résolution de problème basée sur l'utilisation d'expériences passées appelées cas.

Le Raisonnement à Partir de Cas « RàPC » (ou *Case Based Reasoning* : C.B.R.) gère une mémoire pour stoker ces différents cas. Pour résoudre un nouveau problème, il commence par rechercher dans cette mémoire le cas le plus proche de ce problème. Par la suite le système adapte l'ancien cas au problème nouveau pour en déduire une solution à partir de la solution de l'ancien cas [AAM 94].

D'un point de vue cognitif, le raisonnement par cas est à la base de théories psychologiques du comportement humain, en particulier lors de la prise de décision. Or, le recours à des situations passées pour résoudre un problème n'est pas original, car dans notre vie quotidienne, on utilise souvent les connaissances acquises et nos expériences passées pour trouver des solutions à de nouvelles situations [MAL 96].

Dans ce chapitre, nous présentons les principes fondamentaux du RàPC et nous donnons les notions de base concernant le diagnostic médical ainsi que les principaux systèmes RàPC médicaux.

# 2. Approche du raisonnement à partir de cas

Le raisonnement à partir de cas est un des types de raisonnement en IA, dans le domaine de l'apprentissage automatique. Raisonner à partir de cas signifie se remémorer des situations passées, similaires à la situation courante et utiliser ces situations pour aider à résoudre la situation courante. Le raisonnement à partir de cas est une forme de raisonnement par analogie. L'analogie proprement dite recherche les relations de cause à effet dans les situations passées pour les transposer à la situation courante ainsi que les ressemblances entre les situations passées et la situation courante. Le raisonnement à partir de cas recherche seulement les ressemblances ou les relations de proximité entre les situations passées et la situation courante. Le RàPC envisage le raisonnement comme un processus de remémoration d'un petit ensemble de situations concrètes : les cas. Il fonde ses décisions sur la comparaison de la nouvelle situation (cas cible) avec les anciennes (cas sources). Le principe général du

RàPC consiste à traiter un nouveau problème (cas cible) en se remémorant des expériences passées voisines (cas de référence). Ce type de raisonnement repose sur l'hypothèse suivante : si une expérience passée et la nouvelle situation sont suffisamment similaires, alors tout ce qui peut être expliqué ou appliqué à l'expérience passée (base de cas) reste valide si on l'applique à la nouvelle situation qui représente le nouveau problème à résoudre [AAM 94].

### 2.1. Historique

Inspiré par les travaux de Minsky et Schank réalisés à la fin des années 70, Schank [SCH 82] formule pour la première fois le paradigme de raisonnement basé sur les cas. En effet, la théorie développée par Minsky [MIN 75] présente un réseau de nœuds et de relations entre ces nœuds ainsi que la notion de « frame (script, schéma) » qui correspond à une structure remémorée qui doit être adaptée pour correspondre à la réalité d'une nouvelle situation rencontrée. Cependant Schank doute de la flexibilité du raisonnement logique et d'une représentation des connaissances ordinaires sous une forme synthétique de propositions indépendamment vraies. Par conséquent, il reprend ces travaux et suppose que le processus de compréhension correspond à un processus d'explication qui s'applique d'une manière itérative [SCH 82]. D'ailleurs, Schank est considéré comme l'initiateur du terme « Case Based Reasoning ». Il introduit à travers le modèle de « mémoire dynamique » un degré de généralité varié connu sous le nom de « MOPS (Memory Organization Packets) » constituant un réseau dense d'expériences. De plus, l'auteur tente d'opérationnaliser le comportement humain et l'optimiser si possible. Dans ce cadre, Gebhardt et al. [GEB 97] définissent le raisonnement à partir d'expériences comme une façon naturelle de penser caractérisant la réflexion humaine sans doute plus encore que le raisonnement avec des règles.

A la fin des années 80, les recherches dans le domaine du RàPC ont réellement commencé à prendre forme et notamment avec les conférences « DARPA » organisées aux Etats-Unis en 1988 [KOL 88], avant de s'imposer en Europe avec la première conférence Européenne en 1993 à Kaiserslautern [RIC 93], puis avec la première conférence internationale à Lisbonne en 1995 [VEL 95].

### 2.2. Les origines des systèmes RàPC

Les origines du raisonnement à partir de cas sont la psychologie cognitive pour l'étude de la mémoire et le raisonnement par analogie. Le raisonnement à partir de cas complète le raisonnement par analogie par un mécanisme de mémorisation et d'extraction des

expériences. C'est aussi une analogie intra-domaine, conçue pour une tâche bien précise. Par les mécanismes qu'il met en œuvre, notamment ceux liés à l'exploitation de la base de cas, le RàPC entretient aussi des relations avec d'autres formes de raisonnement, comme la classification et la catégorisation [HAT 91].

### 2.2.1. Théorie de la mémoire

Plusieurs théories de la mémoire ont successivement dominé dans les systèmes RàPC. La dernière étant la théorie de la mémoire dynamique de Schanck [SCH 77]. Elle a donné lieu aux premiers systèmes de raisonnement à partir de cas [KOL 93] : la théorie de Lindsay, la théorie de la mémoire épisodique, la mémoire conceptuelle et la théorie de la mémoire dynamique : selon laquelle les processus cognitifs de compréhension, de mémorisation et d'apprentissage utilisent une même structure de mémoire. Cette structure, les "memory organization packets" (ou MOPS) est représentée à l'aide de schémas de représentation de connaissance tels que les graphes conceptuels et les scripts.

### 2.2.2. Le raisonnement par analogie

Par définition, l'analogie désigne « *rapport, similitude partielle d'une chose avec une autre* ». De nombreux exemples d'analogie existent dans notre entourage : l'analogie atome/système solaire, l'analogie courant électrique/cours d'eau. Le raisonnement par analogie consiste à avoir recours à un élément mieux connu pour inférer des informations sur un élément qui l'est moins. Il implique notamment l'évaluation de la ressemblance entre entités et leur mémorisation en vue de leur réutilisation [KOL 93].

#### 2.2.2.1. Finalités du raisonnement par analogie

Le raisonnement par analogie est reconnu être très utilisé par l'être humain. Dans la vie quotidienne, face à une situation donnée l'expérience d'une situation semblable peut s'avérer très utile. En intelligence artificielle, les systèmes mettent en œuvre ce raisonnement à diverses fins : pour la compréhension du langage naturel, la planification, etc.
Le raisonnement par analogie a pour objectif l'inférence d'informations sur une situation (la cible) à partir de la description d'une situation dans laquelle ces informations sont connues (la source).

Les deux composantes essentielles de l'analogie sont la mise en évidence des caractéristiques ou propriétés communes des situations et la détermination des relations intra- ou inter-domaines. En raisonnement par analogie, la ressemblance entre situations est basée souvent sur des critères syntaxiques (le raisonnement à partir de cas au contraire compare plutôt des ensembles de descripteurs) [KOL 93].

Il existe principalement deux contextes d'utilisation du raisonnement par analogie notamment dans le cadre de la résolution de problèmes :

- La résolution du problème s'annonce particulièrement complexe et longue. L'utilisation de la solution d'un problème similaire déjà résolu permet d'accélérer le processus de résolution,

- Dans le domaine considéré, il n'existe pas de théorie qui permet de résoudre le problème posé. L'utilisation d'un problème similaire résolu s'avère être la seule issue.

La première analogie est qualifiée d' «heuristique », et la seconde d'« analogie-recours ».

#### 2.2.2.2. Formalisation et terminologie

Le paradigme d'analogie définit l'analogie comme la mise en œuvre d'un mécanisme de mise en correspondance ou de projection, entre des structures afin de transposer des connaissances d'un univers de base vers un univers cible, en fonction d'un certain point de vue pouvant correspondre à un but à atteindre ou un problème à résoudre [KOL 93].

Ce paradigme manipule deux entités appartenant chacune à un univers pouvant être réorganisé selon le point de vue adopté. L'une des deux entités est utilisée pour inférer des connaissances sur la seconde, elle appartient à l'univers de base, l'autre appartient à l'univers cible. Dans la description des mécanismes du raisonnement par analogie, ces deux entités sont appelées respectivement « base » et « cible».

Un problème d'analogie s'exprime selon l'expression : *« D est à C ce que B est à A. Connaissant A, B et C, que vaut D ?»*. Le principe d'analogie est souvent schématisé par « le carré d'analogie » (Fig.1.1):

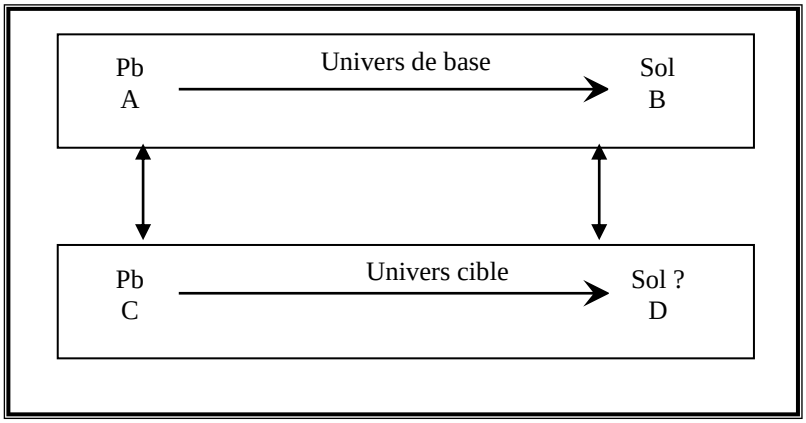



Fig.1.1. Carré d'analogie

Appliqué à la résolution de problème, ce principe a donné lieu à une définition plus précise [SMA 94]: « *La résolution de problèmes par analogie consiste à transférer de la connaissance à partir d'épisodes précédents de résolution, aux nouveaux problèmes qui partagent des aspects significatifs de l'expérience précédente et à utiliser les connaissances transférées pour construire des solutions pour les nouveaux problèmes* ».

Par rapport au carré d'analogie, A et B définissent un épisode de résolution, par un problème (A) et sa solution (B). C représente le problème à résoudre et D la solution recherchée. Carbonell propose deux approches de résolution de problème par analogie [SMA 94] :

- ***Analogie par transformation*** : Elle tente de réutiliser, avec des modifications, une solution précédemment trouvée pour un problème similaire.

Si *d1* représente la différence entre A et C, il s'agit de déterminer *d2*, la différence entre B et D afin d'en déduire D en propageant d2 dans B. Le carré d'analogie devient alors (Fig.1.2) :

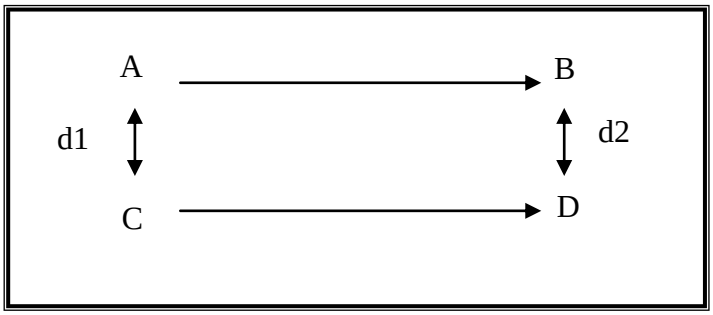



Fig.1.2. Analogie par transformation

- ***Analogie par dérivation*** : Il s'agit d'adapter le processus (ou la dérivation) ayant abouti à la solution d'un problème précédent pour construire la solution d'un problème similaire. Si ***S*** est la similarité entre A et C et ***P*** la méthode de construction de la solution B à partir du problème A. La résolution du problème C se fait alors par le calcul d'une méthode ***p'*** à partir de ***P*** en utilisant ***S*** (***S*** sert à reconnaître les éléments de ***P*** qui sont encore valables dans la nouvelle situation C) (Fig.1.3) :

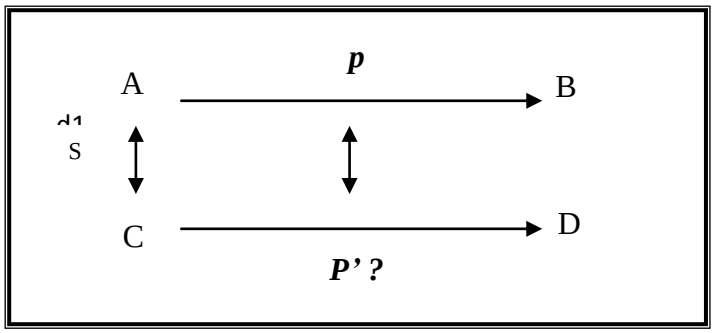



Fig.1.3. Analogie par dérivation

### 2.3. Représentation d'un cas

Un cas dans une base de cas représente différents types de connaissances qui peuvent être stockées selon différents formats de représentations. Un cas est un ensemble de problèmes se note « pb » et la solution de ce problème est alors codifiée « sol (pb) » : *cas=(pb, sol (pb)).*

Un « cas source » est un cas dont on va s'inspirer pour résoudre un nouveau problème et l'on nommera le « cas cible ». La codification est alors la suivante :

Cas-source = (source, sol (source))

Cas-cible = (cible, sol (cible))

Dans chacun de ces types de systèmes RàPC, un cas peut être représenté différemment. Selon les applications du RàPC, un cas, son problème et sa solution, sont décrits par un ensemble de descripteurs. Un descripteur est tout les informations qui nous permettent de décrire le problème. Le descripteur « d » est caractérisé par une paire d=(a,v), où « a » est un attribut défini par un nom et « v » est la valeur qui lui est associée [GEB 97]. Un attribut représente une caractéristique du domaine applicatif qui peut être numérique ou symbolique.

Nous pouvons écrire un cas comme suit :

- $Source = \{d_1^S \dots d_n^S\}$ où $d_i^S$ est un descripteur du problème source.
- $Sol(source) = \{D_1^S \dots D_n^S\}$ où $D_i^S$ est un descripteur de la solution source.
- $Cible = \{d_1^C \dots d_n^C\}$ où $d_i^C$ est un descripteur du problème cible.
- $Sol(cible) = \{D_1^C \dots D_n^C\}$ où $D_i^C$ est un descripteur de la solution cible.

## 3. Composantes d'un système de Raisonnement à Partir de Cas

Un système RàPC est une combinaison de processus et de connaissances ("knowledge containers") qui permettent de préserver et d'exploiter les expériences passées. Pour simplifier la présentation, nous nous appuyons sur le modèle générique présenté dans la figure (Fig.1.4) les principaux processus dont la recherche ("retrieval"), l'adaptation ("reuse"), la maintenance ("retain") et la construction ("authoring"). Les structures de connaissances sont : le vocabulaire d'indexation, la base de cas, les métriques de similarité et les connaissances d'adaptation [HAJ 04].

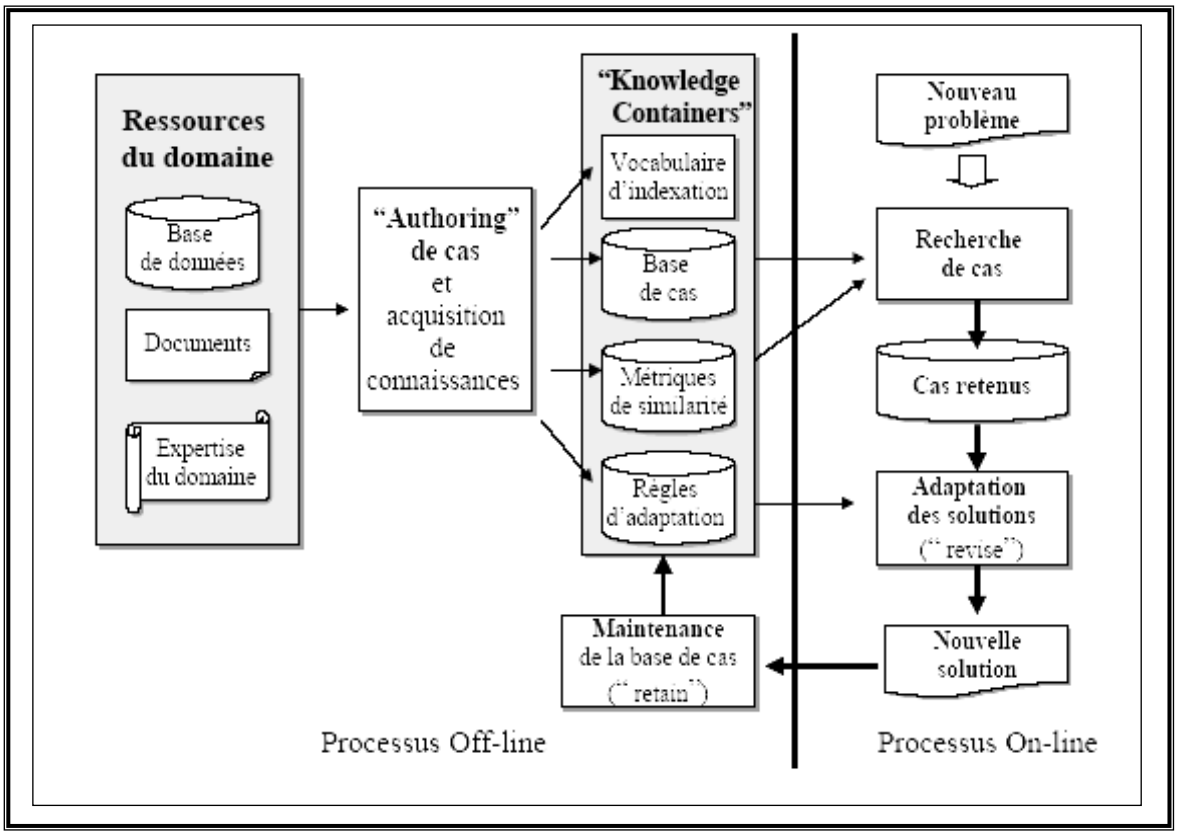



Fig.1.4. Modèle générique d'un système RàPC [AFO 04]

### 3.1. Processus

Le souci fondamental des systèmes RàPC est d'éviter de reproduire les erreurs passées, et de faciliter l'acquisition des connaissances. Le RàPC doit donc se rappeler des cas pertinents de mémoire, puis à partir de la liste des cas retrouvés à l'étape précédente, sélectionner les cas les plus promoteurs en utilisant les mesures de similarités et construire une solution ou une interprétation pour le nouveau cas. Une solution est élaborée en adaptant les anciennes solutions. Ensuite, le RàPC doit tester et critiquer la sortie de l'étape précédente et proposer des contres exemples. Puis il évalue et analyse les résultats dans le monde réel pour enfin mettre à jour la mémoire en stockant et en indexant le nouveau cas [AFO 04] [ARM 09].

Lorsqu'un nouveau cas est à résoudre, il est intégré dans un cas cible où la partie solution est inconnue et doit être apportée par le raisonnement. Les cas sources représentent des expériences passées stockées dans une mémoire. Les cas sources et cibles ont le même formalisme de représentation. L'objectif du raisonnement en cinq phases est de transférer des enseignements pertinents des cas sources pour élaborer la solution du cas cible.

La figure (Fig.1.5) résume clairement le cycle d'un RàPC :

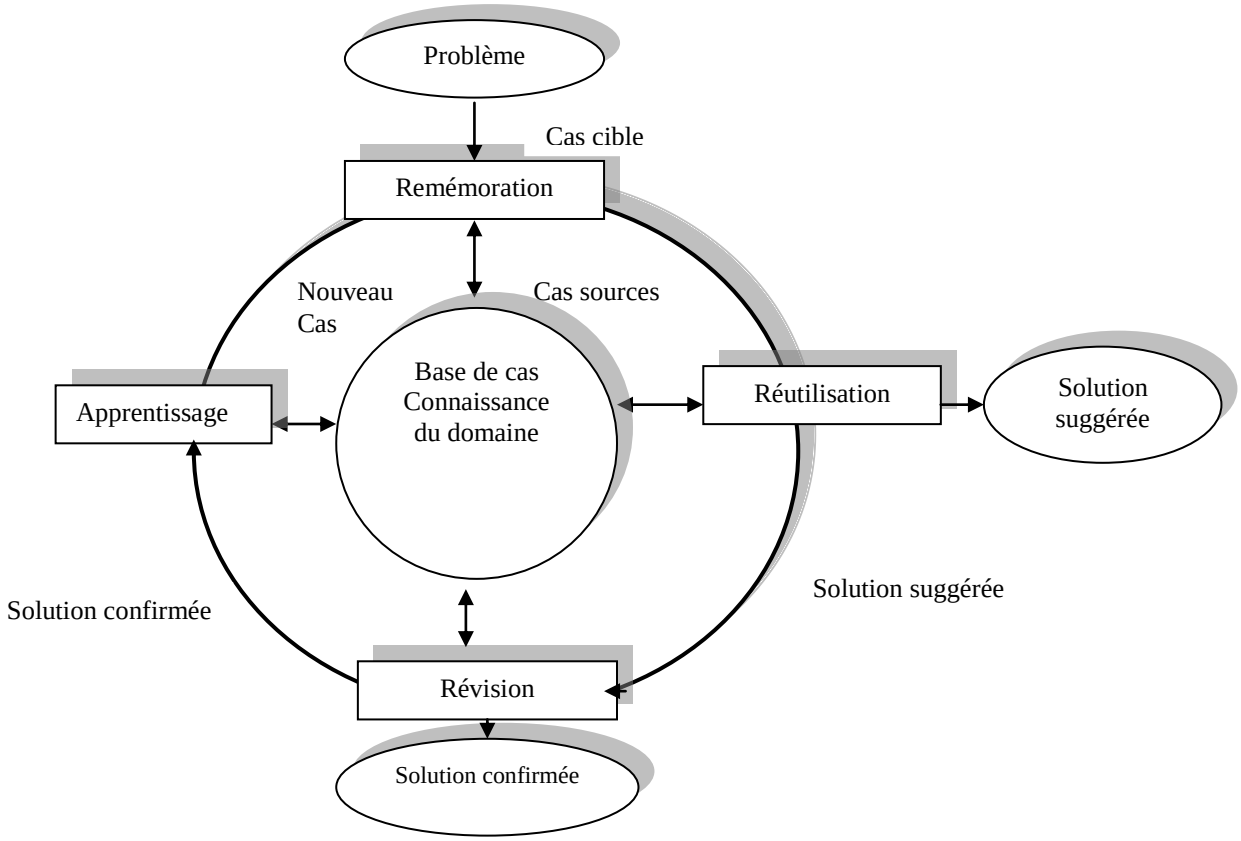



Fig.1.5. Cycle du raisonnement à partir de cas [AFO 04]

### 3.1.1. Remémoration

La phase de remémoration est une étape importante dans un système RàPC. L'une des hypothèses majeures dans le RàPC est de dire que des expériences similaires peuvent guider de futurs raisonnements, des résolutions de problèmes et permettre un apprentissage de la base de cas. La remémoration dépend essentiellement de la représentation de cas, de leur indexation et de leur organisation de la base de cas. Cette phase est donc le processus qui permet de retrouver des cas sources similaires au cas cible. Les cas sources similaires sont des cas qui ont besoin de moins d'adaptation [SOL 04] [HAO 09].

#### 3.1.1.1. Les techniques de recherche des cas similaires

- ***Mesure de similarité***

Ces mesures de similarité cherchent des correspondances entre descripteurs de ces cas qui sont ensuite retrouvés à l'aide d'un algorithme de recherche. L'objectif de ces mesures de similarité est de retrouver le cas de la base de cas similaire au problème actuel au sens qu'il est facilement adaptable à ce nouveau problème. Il existe différentes méthodes pour calculer la similarité ou la dissimilarité. La distance de Minkowski est la plus connue et est souvent utilisée. Il s'agit de calculer la distance relative entre le cas cible et le cas source [FUC 08][PAL 04][ARM 09].

$$Sim(C,S) = \left[\sum_{i=1}^{n} w_i * |C_i - S_i|^p\right]^{1/p}$$

$N$ : est le nombre d'attribut

$C$ : est le cas cible

$C_i$ : est le descripteur de cas cible

$S$ : est le cas source

$S_i$ : est le descripteur du cas source

$W$ : est le poids de l'attribut i

En fonction du paramètre p :

P=1 distance de Manhattan

P=2 distance euclidienne

P=∞ distance de Chebychev (Max$|C_i - S_i|$)

Plusieurs techniques de similarité existent dans la phase de remémoration : *K Plus Proches Voisins (KPPV)*, les *approches inductives*, la *logique floue* ou encore les *réseaux neurones*. Nous décrivons par la suite les deux approches plus utilisées dans ce domaine : *L'approche K Plus Proches Voisins* ainsi que *les approches inductives*.

***L'approche K Plus Proches Voisins***

La technique K Plus Proches Voisins *« KPPV »* (plus connus en Anglais sous le nom *K*-Nearest Neighbors (*K*-NN) [WEI 91]) est la technique la plus utilisée dans les RàPC. Cette technique est établie selon la mesure de Minskowski. Elle est basée sur les poids des descripteurs $w_i$ qui évaluent la similarité entre les cas sources de la base de cas et le cas cible. Le poids permet alors de modifier l'importance d'un descripteur par rapport à un autre descripteur. La partie la plus délicate dans cette technique est de définir de poids des descripteurs. Dans l'équation 1, *w* est le poids du descripteur, *sim* est la fonction de similarité, et $d_i^C$ et $d_i^S$ sont des valeurs du descripteur i dans le cas cible et respectivement dans le cas source.

$$Similarité(C,S) = \frac{\sum_{i=1}^{n} w_i \, x \, sim\left(d_i^C, d_i^S\right)}{\sum_{i=1}^{n} w_i} \qquad (1)$$

***Les approches inductives***

Les approches inductives déterminent quels sont les descripteurs qui apportent un meilleur effet pour une discrimination des cas. Les algorithmes d'Induction, tels que ID3 [QUI 86] et CART (Classification And Regression Trees), permettent de construire des arbres de décision à partir des cas de la base de cas. Ces algorithmes d'induction divisent les cas en grappes (clusters). Chaque cluster contient des cas similaires. Les techniques d'Induction sont utilisées majoritairement comme des classificateurs pour regrouper des cas similaires. Ils déterminent quels descripteurs sont à même de mieux distinguer les cas et de générer une structure de l'arbre de décision pour organiser le cas de la base de cas.

### 3.1.2. Adaptation ou réutilisation

Suite à la sélection de cas lors de la phase de recherche, le système RàPC aide l'usager à modifier et à réutiliser les solutions de ces cas pour résoudre son problème courant. En général, on retrouve deux approches d'adaptation de cas [BUI 04] [DAQ 04] [HAO 09]:

- ✓ *l'approche transformationnelle (ou structurelle)* : on obtient une nouvelle solution en modifiant des solutions antécédentes et en les réorientant afin de satisfaire le nouveau problème.

- ✓ *l'approche générative (ou dérivationnelle) :* on garde, pour chaque cas passé, une trace des étapes qui ont permis de générer la solution. Pour un nouveau problème, une nouvelle solution est générée en appliquant l'une de ces d'étapes.

Peu de systèmes RàPC font de l'adaptation complètement automatique. Pour la plupart des systèmes, une intervention humaine est nécessaire pour générer partiellement ou complètement une solution à partir d'exemples. Le degré d'intervention humaine dépend des bénéfices en terme de qualité de solution que peut apporter l'automatisation de la phase d'adaptation [HAM 89].

### 3.1.3. Maintenance ou révision

Durant le cycle de vie d'un système CBR, les concepteurs doivent préconiser certaines stratégies pour intégrer de nouvelles solutions dans la base de cas et pour modifier les structures du système RàPC pour en optimiser les performances. Une stratégie simple est d'insérer tout nouveau cas dans la base. Mais d'autres stratégies visent à apporter des modifications à la structuration de la base de cas (e.g. indexation) pour en faciliter l'exploitation. On peut également altérer les cas en modifiant leurs attributs et leur importance relative. Cet aspect de recherche est actuellement l'un des plus actifs du domaine RàPC [LEA 01] [SOL 04].

Cette étape a pour but la validation de la solution produite. La confrontation de cette sol ution au monde réel détecte les erreurs des deux étapes précédentes. Elle est généralement réal isée dans une boucle évaluation / correction.

### 3.1.4. Construction ou apprentissage

Ce processus, en amont des activités de résolution de problèmes du système RàPC, sous-tend la structuration initiale de la base de cas et des autres connaissances du système à partir de différentes ressources tels des documents, bases de données ou transcriptions d'interviews avec des praticiens du domaine. Ce processus, souvent effectué manuellement par le concepteur du système, se prête moins bien à l'automatisation car il nécessite une connaissance du cadre applicatif pour guider, entre autre, la sélection du vocabulaire d'indexation et la définition des métriques de similarités [AAM 98].

### 3.2. Connaissances dans un système RàPC

Richter [RIC 98] définit les systèmes de RàPC comme des systèmes à base de connaissances (SBC). Ces systèmes exploitent quatre catégories de connaissances ("knowledge containers") distinctes à savoir :

- ✓ *vocabulaire d'indexation* : un ensemble d'attributs ou de traits ("features") qui caractérisent la description de problèmes et de solutions du domaine. Ces attributs sont utilisés pour construire la base de cas et jouent un rôle important lors de la phase de recherche.
- ✓ *base de cas* : l'ensemble des expériences structurées qui seront exploitées par les phases de recherche, d'adaptation et de maintenance.
- ✓ *mesures de similarité* : des fonctions pour évaluer la similarité entre deux ou plusieurs cas. Ces mesures sont définies en fonction des traits et sont utilisées pour la recherche dans la base de cas.
- ✓ *connaissances d'adaptation :* des heuristiques du domaine, habituellement sous forme de règle traits permettant de modifier les solutions et d'évaluer leur applicabilité à de nouvelles situations.

## 4. Organisation de la mémoire de cas

Les processus de mémorisation et de remémoration sont fortement liés à la façon d'organiser les cas dans la base. Dans cette section, nous allons aborder différentes méthodes d'organisation des cas en mémoire. Elles se résument en deux catégories principales [KOL 93] [AAM 94]:

- ➢ *La mémoire plate* : il s'agit de mémoriser tous les cas dans une liste séquentielle. Nous détaillons ce type de mémoire dans la section suivante.
- ➢ *La mémoire hiérarchique* : lorsque la mémoire des cas est large, il y a une nécessité d'organiser les cas hiérarchiquement. Ceci permet de simplifier la remémoration. Par la suite, nous présentons deux approches pour l'organisation hiérarchique des cas en mémoire : *les réseaux à trait partagés* et *les arbres de discriminations*.

### 4.1. Organisation plate

Les cas sont stockés séquentiellement dans une simple liste ou fichier, c'est la plus simple structure à imaginer pour une mémoire. Les avantages d'une telle structure est que pendant la remémoration, tous les cas existant dans la mémoire sont testés, ceci garantie une remémoration précise qui dépend de la quantité de la fonction d'appariement. De plus, la mémorisation n'est pas coûteuse, il suffit d'ajouter le nouveau cas à la fin du fichier.

L'inconvénient majeur est le temps de remémoration qui augmente linéairement avec la taille de la mémoire. Il existe plusieurs variantes pour l'organisation en mémoire plate [MAL 96]:

- *Indexation superficielle :* l'indexation se fait dans un seul niveau, chaque descripteur (attribut-valeur) choisi comme étiquette pointe vers les cas correspondants (qui contiennent ce descripteur dans leur représentation).

Pendant la phase de remémoration, les cas qui sont pointés par ce descripteur sont sélectionnés, et ensuite la fonction d'appariement est appliquée à ce groupe des cas, et non à tous les cas. Ceci fonctionne très bien quand ces descripteurs permettent d'extraire un petit nombre de cas quand les étiquettes sont suffisamment descriptives.

- *Partitionnement de la mémoire :* le système doit reconnaître à quelles partitions une nouvelle situation appartient. La fonction d'appariement est appliquée seulement aux cas appartenant aux partitions sélectionnées.

- *Extraction parallèle :* la fonction d'appariement est d'appliquée d'une manière parallèle à tous les cas qui existent dans la mémoire.

### 4.2. Réseaux à caractéristiques partagées

Les réseaux à caractéristiques partagées regroupent les cas présentant des similarités dans un même cluster. Les hiérarchies sont formées lorsque les clusters sont subdivisés en sous clusters. Cette organisation offre l'avantage de mieux partitionner la base de cas et rend la recherche plus efficace. Mais quelques inconvénients, telle que la complexité lors de l'ajout d'un cas, la difficulté de maintenir l'optimalité du réseau lors de l'ajout. Un espace supplémentaire est requis pour l'organisation.

Plusieurs réseaux avec des priorités différentes seraient nécessaires pour augmenter la précision de la recherche.

### 4.3. Réseaux de discrimination

Le regroupement effectué dans les réseaux à caractéristiques partagées conduit à une discrimination en second lieu. Dans ces réseaux, chaque nœud interne est une question qui départage selon la réponse des cas de la base. Les questions les plus importantes sont posées en premier. Ce type d'organisation, en plus des avantages des réseaux à caractéristiques partagées, rend intuitive la compréhension de la connexion entre indices et l'organisation du réseau. Les inconvénients sont ceux des réseaux à caractéristiques partagées, plus le problème de traitement des informations manquantes.

### 4.4. Réseaux redondants de discrimination

Les réseaux redondants de discrimination fournissent une réponse au problème des informations manquantes. Ils organisent les cas en utilisant différents réseaux de discriminations, chacun avec un ordre différent des questions. Une recherche se fait en parallèle sur les différents réseaux. Si dans l'un des réseaux une question n'a pas de réponse, on y abandonne la recherche. Au moins, l'un des réseaux retrouvera le cas qui s'apparie s'il existe. L'inconvénient majeur d'une telle organisation est la complexité de sa mise en œuvre.

### 4.5. Exemple de modèles hybrides de mémoire de cas

#### 4.5.1. Le modèle PROBIS : un modèle hybride de mémoire

PROBIS (Prototype-Based Indexing System) est un modèle de mémoire hybride proposé par Malek en 1996 [MAL 96] intégrant un réseau incrémental à base de prototypes et une mémoire plate partitionnée en plusieurs groupes. Ce modèle complet permet de combiner plusieurs avantages : une remémoration efficace et précise en même temps, une mémorisation simple et un traitement des cas atypiques et frontières. Ce modèle forme une mémoire à deux niveaux de hiérarchie: le bas niveau qui comprend une mémoire plate contenant des cas organisés en groupes ou chaque groupe contient un ensemble de cas similaires; le haut niveau qui contient des prototypes, chaque prototype représente un groupe de cas du bas niveau. L'ensemble de ces prototypes forme un système d'indexation pour les différents groupes dans la mémoire plate. L'utilisation d'un réseau incrémental à base de prototypes permet de

construire les prototypes représentatifs des différents groupes et d'organiser les cas dans les différents groupes.

### 4.5.2. Réseaux de recherche de cas : CRN

Le principe des CRNs (Case Retrieval Nets) est inspiré des réseaux de neurones et des modèles de mémoires associatives. L'idée de [NOU 04] est que le processus de rappel d'un cas ne se fait pas en parcourant un chemin dans une arborescence mais plutôt de façon reconstructive en récupérant graduellement les entités d'information constituants le cas. Les connaissances de base dans les CRNs sont les entités d'information (IEs). Un cas est un ensemble de ces entités. Une mémoire de cas est alors : un réseau de nœuds correspondant aux IEs du domaine ainsi que de nœuds additionnels dénotant le cas. Les nœuds IEs sont connectés par des arcs de *similarité* et les nœuds cas sont accessibles à partir des IEs les constituants à travers des arcs de *pertinence***.** Différents degrés de similarité et de pertinence peuvent être exprimés en faisant varier le poids des arcs.

## 5. Dignostic 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.

### 5.1. 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. 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 traités 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 [ALS 12].

### 5.2. Système d'aide au diagnostic médical

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 :

« *Software designed to be a direct aid to clinical decision-making, in which the characteristics of an individual patient are matched to a computerized clinical knowledge base and patient specific assessments or recommendations are then presented to the clinician or the patient for a decision* ».

Kawamoto et *al.* [KAW 05] ont défini le système d'aide au diagnostic comme suit :

« *We defined a clinical decision support system as any electronic or non-electronic system designed to aid directly in clinical decision making, in which characteristics of individual patients are used to generate patient-specific assessments or recommendations that are then presented to clinicians for consideration* ».

Ces différentes définitions confirment le fait que l'aide (i.e. informations obtenues par le système) fournie au médecin dans son processus de diagnostic, peut prendre plusieurs formes (i.e. cas similaires déjà diagnostiqués, diagnostics potentiels, etc.). En effet, les systèmes d'information, les bases de données et les dossiers informatisés facilitent la prise de décision en améliorant l'accès aux données pertinentes et leur mise en perspective. Néanmoins, il ne s'agit que d'une aide indirecte présentant des faits sur lesquels le décideur doit appliquer un raisonnement. Les systèmes d'aide à la décision ont l'ambition d'assister le médecin, en

remplaçant ou en reproduisant le raisonnement humain. 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 à partir de cas. Par la suite, nous présentons les principaux systèmes de diagnostic médical par le RàPC.

## 6. Les principaux systèmes RàPC en diagnostic médical

En médecine, le RàPC a été utilisé à de nombreuses reprises pour la conception d'outils d'aide à la décision. En effet, la médecine est un domaine à fort potentiel pour ce qui est de la prise de décision, les cas illustrant des exemples de prises de décisions médicales sont assez répandus. Il existe un lien intense entre la médecine et le raisonnement à partir de cas [HEI 99]. Ce dernier étant un paradigme bien adapté pour le développement d'un système d'aide à la prise de décision médicale. Certainement, un des dispositifs intuitivement attrayants des systèmes RàPC dans la médecine est que le concepts du *patient* et de la *maladie* se prêtent naturellement à une représentation de cas [NIL 04]. De nombreux autres exemples de systèmes de RàPC médicaux sont proposés dans les ouvrages de référence sur le RàPC dans [WAT 97] [KOL 93]. Dans la suite, nous décrivons quelques systèmes RàPC de diagnostic médical les plus connus.

### 6.1. Système CASEY

Le système d'aide au diagnostic « CASEY» est réalisé par Koton en 1988 [KOT 88]. L'entrée de ce système est une description du nouveau patient et qui peut être des signes normaux, des signes présents et/ou des symptômes. La sortie est une explication causale de la maladie. Cette explication relie les symptômes à l'état interne. CASEY procède au diagnostic des patients en appliquant un appariement basé modèle et une adaptation heuristique. Sa base de cas contient 25 cas tous diagnostiqués par le programme de détection des problèmes cardiaques (HFP pour Heart Failure Program) qui est un raisonneur basé modèle (Fig.1.6).

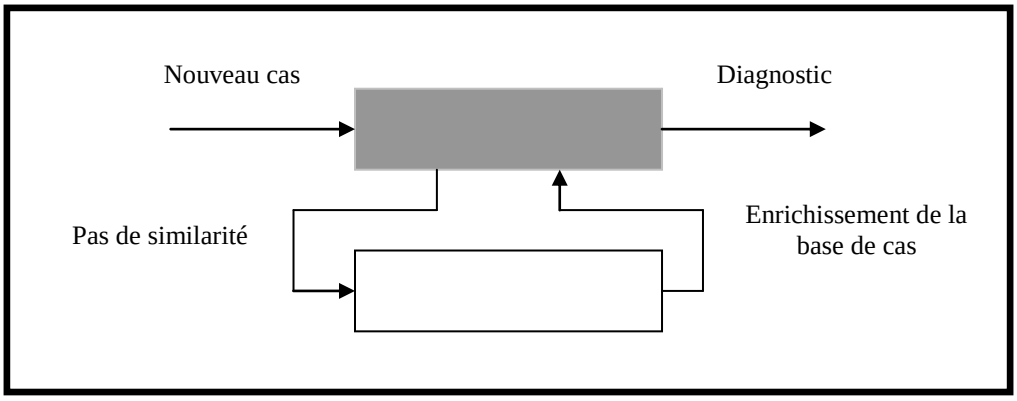



Fig.1.6. Coopération CASEY/ HFP [KOL 93]

Lorsqu'un nouveau patient se présente, CASEY cherche un cas similaire. S'il en trouve, il l'utilise pour le diagnostic du nouveau patient. S'il n'en trouve pas, il soumet le nouveau cas au HFP qui le diagnostique et retourne les résultats à CASEY pour une utilisation future (enrichissement de la base). CASEY utilise un processus à deux étapes pour diagnostiquer un nouveau cas. Il procède à une recherche des cas similaires en utilisant des règles d'évidence pour déterminer parmi les cas s'appariant partiellement quels sont ceux suffisamment similaires pour proposer un diagnostic précis. Puis, il applique des règles de réparation pour adapter l'ancien diagnostic à la nouvelle situation. CASEY utilise, pour son adaptation, un ensemble de règles de réparation basées modèle indépendantes du domaine. Chaque règle de réparation est associée à une ou plusieurs règles d'évidence. CASEY garde la trace de la règle d'évidence appliquée et active la règle de réparation correspondante lors de l'adaptation.

**6.2. Système PROTOS**

PROTOS [BAR 88] est un système de RàPC pour l'aide au diagnostic des malentendants. Le cas est représenté par un vecteur de caractéristiques. Il n'y a aucun modèle utilisé et la phase de remémoration est basée sur l'algorithme des k plus proches voisins. Protos ne dispose pas de phase d'adaptation. Si la solution proposée par le système n'est pas satisfaisante alors une phase d'apprentissage s'enclenche. De ce fait, le système commence à rechercher une autre solution sinon il doit accepter une solution proposée par l'utilisateur. Dans ce cas, l'utilisateur est obligé de définir les termes dans le système qui sont inconnus de PROTOS, en décrivant leur relation avec les termes existants. Cette phase permet de construire la connaissance générale du système. Si le problème a été résolu tout de suite, le système renforce (par pondération) les caractéristiques ayant permit la résolution. Parallèlement, si le problème n'est résolu qu'en seconde étape, PROTOS affaiblit les

caractéristiques fautives. L'originalité de PROTOS tient au fait que son apprentissage repose sur l'utilisateur.

### 6.3. Système IDEM

IDEM (Images et Diagnostic par l'Exemple en Médecine) est un système développé en France par le service d'informatique médicale de l'Hôpital Broussais à Paris. Ce système est développé par Bozec [BOZ 98] pour l'aide au diagnostic en imagerie médicale dans le domaine de l'anatomie pathologique mammaire. Les trois fonctionnalités principales du système sont :

- ✓ Se familiariser avec le système.
- ✓ Naviguer dans un ATLAS de cas de pathologie tumorale mammaire.
- ✓ Interroger la base de cas.

Les cas sont des exemples d'interprétations d'images macroscopiques et microscopiques. Pour modéliser ces cas, les auteurs se sont basés sur des comptes-rendus d'experts, et en ont obtenu une représentation composite. Un cas est alors composé d'une arborescence de «zones», elles-mêmes composées d'un ensemble de couples d'<attribut : valeur>. La mesure de similarité préalablement définie, opère à trois niveaux: au niveau de l'attribut, au niveau de la zone et enfin au niveau du cas. Au niveau de la zone et du cas, la mesure de similarité intègre une ressemblance de surface et une ressemblance de structure. La mesure globale de ressemblance est calculée comme la moyenne arithmétique d'une similarité de surface et d'une similarité de structure.

Le tableau (Tab.1.1) résume les principaux systèmes RàPC médicaux réalisés lors de ces vingt dernières années ainsi que le contexte médical de chaque système.

| Références | Système de RàPC | Domaine d'application (contexte) |
|---|---|---|
| [TUR 88] | MEDIC | Problèmes pulmonaires |
| [KOT 88] | CASEY | Problèmes cardiaques |
| [BAR 88] | PROTOS | Aide au diagnostic des malentendants |
| [KAH 93] | ISIS | La radiologie |
| [BRA 93] | FLORENCE | Diagnostic, pronostic et prescription |
| [BIC 94] | MNAOMIA | Troubles du comportement alimentaire |
| [OPI 95] | MERSY | La santé des travailleurs de la région rurale |
| [JUR 98] | TA3-IVE | Des plans pour les traitements des cas de fertilisation in vitro |
| [BOZ 98] | IDEM | Anatomie pathologique mammaire |
| [HAD 97] [GIE 98] | ProtoISIS, MacRad, et SCINA | L'interprétation des images médicales |
| [BRE 00] | CASMIR | Cancer du sein |
| [BAL 03] | Ultranet | Les malformations foetales |
| [BIC 03] | SFDA | La grippe porcine |
| [SCH 05] | TeComMed | Des prévisions sur les épidémies de grippe |
| [SCH 07] | ICONS | Des conseils dans le traitement antibiotique |
| [SCH 09] | ISOR | L'explication et l'interprétation de situations exceptionnelles |
| [BEG 09] | IPOS | Stress |

Tab.1.1 Tableau récapitulatif des principaux systèmes de diagnostic médical par rapport au contexte

## 7. Discussion : vers un model probabiliste de mémoire de cas

Nous avons parcouru dans ce chapitre (session.4), les différents aspects liés à la mémoire dans un système de raisonnement à partir de cas. Nous avons montré l'importance de la mémoire des cas au sein d'un tel système pendant les processus d'apprentissage (ou de mémorisation), de remémoration (ou de rappel) et d'adaptation (réutilisation). Nous avons aussi évoqué les problèmes liés à l'organisation des cas en mémoire. Par ailleurs, nous allons décrire une comparaison entre les différentes organisations de mémoire de cas.

Les modèles de mémoire parcourus dans ce chapitre montrent que deux grandes approches pour l'organisation de mémoire sont distinguées : l'approche de la mémoire plate et l'approche fondée sur l'organisation hiérarchique de la mémoire.

Dans une mémoire plate, les cas sont stockés d'une manière séquentielle dans un fichier. La remémoration consiste alors à appliquer une fonction d'appariement sur chacun

des cas. L'avantage de cette approche est qu'elle garantie une bonne précision grâce à l'examen de toute la base de cas. De plus, la mémorisation d'un nouveau cas est facile. Elle se traduit par l'ajout de ce cas à la fin de la mémoire. En revanche, le temps de la remémoration augmente linéairement avec la taille de la base de cas. Et l'adaptation du cas similaire dépend essentiellement de l'algorithme d'adaptation.

L'organisation hiérarchique de la mémoire consiste à organiser les cas dans une mémoire avec plusieurs niveaux de hiérarchie. Cette méthode rend la remémoration plus efficace en terme de temps de calcul, mais complique l'apprentissage des nouveaux cas à cause de la structure compliquée de la mémoire. La phase d'adaptation sera une tâche difficile à réaliser.

Le modèle CRN de [NOU 04] est une structure dynamique qui facilite la recherche des cas similaires au cas en entrée. Il simplifie la maintenance de la base de cas. Pour insérer un nouveau cas, il est nécessaire d'ajouter les liens de relevance entre le nœud identificateur du nouveau cas et les entités d'informations existantes qui le détermine. L'adaptation a été améliorée grâce au critère d'adaptabilité.

Le modèle PROBIS de [MAL 96] a permis de combiner les avantages de l'utilisation d'une mémoire hiérarchique (efficacité de remémoration) sans perdre l'avantage de l'utilisation d'une mémoire plate (précision de réponse), il a également contribué à l'augmentation de l'efficacité du système en terme de temps de réponse grâce au traitement parallèle et a enfin simplifié le processus de l'apprentissage en terme de complexité de calcul en comparaison avec les modèles de mémoire hiérarchique.

Le tableau suivant (Tab.1.2) donne une comparaison entre les différentes approches de l'organisation de la base de cas selon des critères liés à la remémoration, à la mémorisation et à l'adaptation.

| | Remémoration | | Mémorisation | Adaptation |
|---|---|---|---|---|
| **Approche** | **Efficacité** | **Précision** | **Simplicité** | **Précision** |
| **Organisation plate** | - | + | ++ | ± |
| **Organisation hiérarchique** | + | - | - | - |
| **Réseaux Neurones** | ++ | - | + | - |
| **Modèle PROBIS** | ++ | + | + | + |
| **Modèle CRN** | ++ | - | + | + |

Tab.1.2. Comparaison entre les différentes organisations de base de cas

*+ : bon, ++ : très bon, - : moyen*

Ces critères sont :

- ***L'efficacité de la remémoration :*** il s'agit de l'efficacité du rappel en terme de temps de calcul. Il est évident que les réseaux sont les plus efficaces grâce à leur capacité du traitement parallèle, ensuite viennent les organisations hiérarchiques, les organisations hybrides et finalement l'organisation plate.

- ***La précision de la remémoration :*** il s'agit de la précision avec laquelle l'algorithme de remémoration extrait les cas les plus proches du problème traité. Il est évident que la précision est maximale quand les cas dans la base sont visités : c'est le cas de l'organisation plate. Dans l'organisation neuronale, cette précision n'est pas toujours garantie car les réseaux peuvent oublier plusieurs exemples au cours de l'apprentissage. Enfin, l'organisation des bases hiérarchiques ne garantissent pas la visite de tous les cas mais juste d'un sous-ensemble.

- ***La simplicité de la mémorisation :*** il s'agit de la complexité du processus d'apprentissage (ajout des nouveaux cas à la base). Il est évident que dans l'organisation plate, l'ajout d'un nouveau cas est le plus simple. En générale dans les réseaux, l'apprentissage est un processus simple aussi tandis que dans les cases hiérarchiques, les algorithmes de mise à jour sont très compliqués.

- ***La précision :*** représente le taux de bons classements parmi les classements proposés par le système.

## 8. Nos choix et notre démarche

Nos choix et nos démarches dépendent de l'organisation de la mémoire de cas choisi. De ce fait, la base de cas est une collection de cas de résolution du même problème. Elle tient une place primordiale dans la mise en place d'un RàPC de qualité. Le problème de la représentation des cas commence toujours par le choix d'informations que l'on veut stoker, pour trouver la structure correspondante, l'organisation des cas et l'indexation [KOL 93][AAM 94]. Dans notre problème de diagnostic, les observations sur lesquelles se basent la décision médicale sont imparfaites, de grande qualité et incertaines dans le processus décisionnel, qu'il soit diagnostic, thérapeutique ou pronostique, c'est donc un processus sous incertitude. C'est pour cette raison, que pour diagnostiquer des pathologies hépatiques, il nous parait nécessaire de modéliser les connaissances par un réseau bayésien.

Les réseaux bayésiens sont des outils efficaces de représentation des connaissances incertaines et ils décrivent hiérarchiquement toutes les étapes du raisonnement du médecin. Nous allons étudier dans le quatrième chapitre la possibilité d'utiliser des méthodes statistiques pour modéliser la mémoire de cas dans le RàPC. Nous nous intéressons à la modélisation de la base de cas par un réseau bayésien et la description des deux phases : la phase de la remémoration et la phase d'adaptation afin d'améliorer les performances du RàPC.

## 9. Conclusion

Le raisonnement à partir de cas a été choisi comme méthode de résolution de problèmes dans notre système d'aide au diagnostic. Nous avons introduit des principes fondamentaux d'un système de raisonnement à partir de cas et un état de l'art des méthodes utilisées dans le cycle de RàPC. Cela nous a permis de poser les jalons pour comparer les différents systèmes de RàPC appliqués au diagnostic médical. Nous avons confirmé notre démarche de modélisation des connaissances présentée au troisième chapitre par une mise en parallèle entre le cycle du RàPC et celui de réseaux bayésiens.

Notons que la mémoire de cas joue un rôle très important dans le RàPC, plus elle s'accroît, et plus les temps de calcul sont longs. C'est pourquoi plusieurs techniques d'organisation de la mémoire ont été proposées. Et comme le domaine de notre travail est la médecine où l'incertitude est un compagnon permanent de l'action médicale. Pour cela, nous avons choisi les réseaux bayésiens comme un modèle d'organisation de la base de cas.

Le chapitre suivant sera alors consacré à la présentation des réseaux bayésiens.

# Chapitre 2

# Les Réseaux Bayésiens

## Sommaire

# 1. Introduction

Un diagnostic médical est le résultat du raisonnement d'un médecin, décision très souvent prise à partir d'informations incertaines et/ou incomplètes. De nombreuses techniques d'intelligence artificielle ont été appliquées pour essayer de modéliser ce raisonnement. Citons, par exemple, des systèmes à base de règles comme MYCIN et Internist-1/QMR (Quick Medical Reference) [KIN 67] [TUR 88].

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

Les réseaux bayésiens constituent une technique d'acquisition, de représentation et de manipulation de connaissance. Toute application mettant en œuvre des connaissances peut relever de l'utilisation des réseaux bayésiens, qu'il s'agisse de formaliser la connaissance d'experts, d'extraire la connaissance contenue dans des bases de données, ou d'utiliser le plus rationnellement possible l'un ou l'autre type de connaissance. Les réseaux bayésiens sont utilisés pour leur capacité d'effectuer des inférences dans un contexte d'incertitude, en quelque sorte comme alternative aux systèmes experts. Comme ils sont utilisés pour leurs algorithmes d'apprentissage et d'inférence, comme alternative aux autres méthodes de modélisation quantitative, en les considérants comme des modèles de régression.

L'utilisation des réseaux bayésiens pose un certain nombre de questions méthodologiques [LER 02] :

. Comment choisir la structure du réseau bayésien ?
. Comment estimer les probabilités du réseau ?
. Comment prendre en compte les données incomplètes ou les variables latentes ?
. Comment faire de l'inférence, exemple: calculer la probabilité de telle ou telle maladie sachant certains symptômes ?

Dans ce chapitre nous allons définir les réseaux bayésiens. Nous présentons les principales définitions, les notions d'indépendance conditionnelle, les principaux théorèmes et les étapes de la construction des réseaux bayésiens. Ensuite, nous élaborons les deux algorithmes d'inférences exactes à savoir : l'algorithme Pearl et l'algorithme de Jensen

Lauritzen Olesen (JLO). Ce chapitre donne une idée détaillée sur les réseaux bayésiens, il aborde les points pertinents pour comprendre cette méthode.

## 2. Définition

Un réseau bayésien est un système représentant la connaissance et permettant de calculer des probabilités conditionnelles apportant des solutions à différentes sortes de problématiques. La structure de ce type de réseau est simple : un **graphe** dans lequel les **nœuds** représentent des variables aléatoires, et les **arcs** (le graphe est donc orienté) reliant ces dernières sont rattachées à des probabilités conditionnelles. Notons que le graphe est **acyclique** : il ne contient pas de boucle. Les arcs représentent des relations entre variables qui sont soit déterministes, soit probabilistes. Ainsi, l'observation d'une ou plusieurs causes n'entraîne pas systématiquement l'effet ou les effets qui en dépendent, mais modifie seulement la probabilité de les observer. L'intérêt particulier des réseaux bayésiens est de tenir compte simultanément de connaissances a priori d'experts (dans le graphe) et de l'expérience contenue dans les données.

Les domaines d'utilisation principaux sont: diagnostic (médical et industriel), analyse de risques, détection de spams, datamining, détection de fraudes, exploitation du retour d'expérience, modélisation et simulation de systèmes complexes, détection d'intrusions, TextMining, analyse de BioPuces, analyse de trajectoires de santé. En bref, un réseau bayésien est un modèle probabiliste graphique permettant d'acquérir, de capitaliser et d'exploiter des **connaissances**, né du besoin de créer des systèmes experts à base de probabilités [MIL 07].

## 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 [Ω,p] tel que :

$$p(V1, V2, \dots, Vn) = \prod_{i=1}^{n} p(Vi|C(Vi)) \qquad (1)$$

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

## 4. 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.

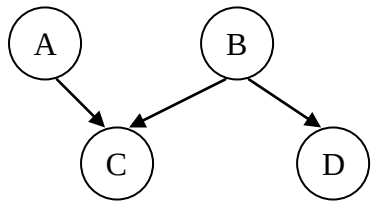
### 4.1. Exemple

Nous allons commencer par l'exemple suivant [PEA 88][FIN 96] :

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

La représentation graphique du modèle causal utilisé est dans la figure (Fig.2.1). Cette figure représente un réseau bayésien simple contenant quatre variables binaires, et on peut écrire aussi :

$$P(A, B, C, D)=P(A).P(B).P(C|A,B).P(D|B) \qquad (2)$$




| | |
|---|---|
| A | Arroseur en marche |
| B | Il a plu pendant la nuit |
| C | Herbe du jardin humide |
| D | Herbe du jardin voisin humide |

P(A=V)=0.4, P(A=F)=0.6, P(B=F)=0.7, P(B=F)=0.3

| A | B | P(C=F) | P(C=V) |
|---|---|---|---|
| V | V | 0.3 | 0.7 |
| V | F | | |
| F | V | 0.05 | 0.95 |
| F | F | | |
| | | 0.4 | 0.6 |
| | | 0.78 | 0.22 |

| B | P(D=V) | P(D=F) |
|---|---|---|
| F | 0.5 | 0.5 |
| V | | |
| | 0.58 | 0.42 |

Fig.2.1. Représentation graphique du modèle

### 4.2. Circulation de l'information

La représentation graphique la plus intuitive de l'influence d'un événement est de relier la cause à l'effet par une flèche orientée. S'il existe une relation causale de A vers B, toute information sur A peut modifier la connaissance sur B, et réciproquement, toute information sur B peut modifier la connaissance sur A [LER 99].

En présence d'un graphe plus complexe, il est essentiel de conserver à l'esprit que l'information ne circule pas seulement dans le sens des flèches (Fig.2.2).

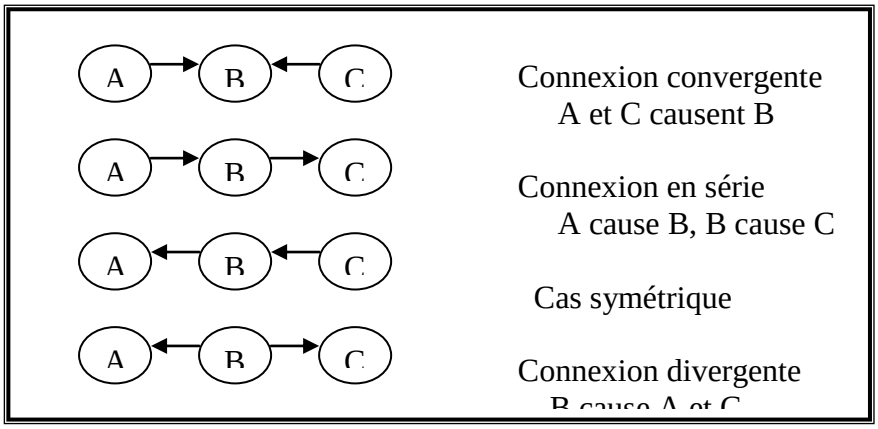



Fig.2.2. Modes de connexion

Dans un graphe on peut rencontrer plusieurs modes de connexion entre les nœuds comme le montre la (Fig.2.3). Nous allons étudier comment l'information circule au sein d'un graphe causal. Dans l'exemple précédent, on voit que l'information a circulé uniquement dans le sens

Effet → cause, par exemple la connaissance de D (herbe du jardin humide) renforce la croyance de la cause B (il a plu). De plus cet exemple nous montre que l'information peut suivre des chemins à première vue contre-intuitifs lorsqu'elle se propage dans un réseau de causalités.

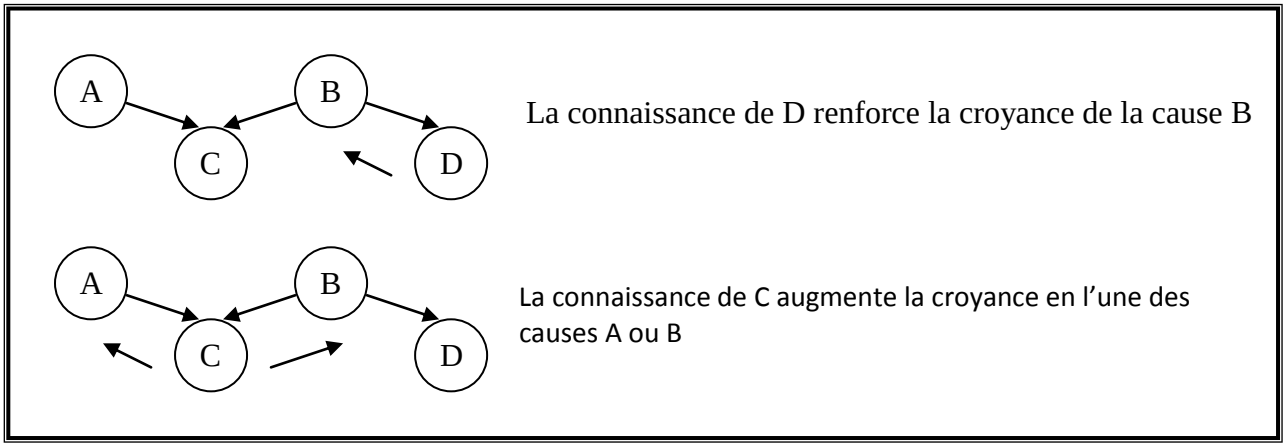



Fig.2.3. Circulation de l'information dans le graphe de la Fig .2.1

### 4.3. Définitions et propriétés

Dans ce paragraphe, nous présentons les définitions relatives aux réseaux bayésiens [FIN 96]:

**Définition 1 :**

Soient

- un graphe acyclique orienté G=(V,A), où V est l'ensemble des nœuds de G et A l'ensemble des arcs de G.
- un espace probabilisé fini (Ω ,Z,P), n variables aléatoires ($X_i$), $1 \leq i \leq n$.

(G, P) est un **réseau bayésien** si et seulement si

- il existe une bijection entre les nœuds du graphe G et les variables ($X_i$).
- la propriété suivante (appelée propriété de factorisation), est vérifiée

$$P(X_1, X_2, \dots, X_n) = \prod_{1 \leq i \leq n} P(X_i | C(X_i)) \qquad (3)$$

où $C(X_i)$ est l'ensemble des causes (parents) de $X_i$ dans le graphe G.

### 4.3.1. Indépendance conditionnelle

**Définition 2 :**

Soient deux variables aléatoires X et Y. on dit que X et Y sont **indépendantes conditionnellement** à Z et on note $X \perp Y|Z$, si l'une des propriétés équivalentes suivantes est vérifiée :

$$P(X|Z,Y) = P(X|Z) \quad (4)$$

$$P(X,Y|Z) = P(X|Z).P(Y|Z) \quad (5)$$

**Propriété 1 :**

Soient X,Y, Z, W des variables aléatoires, on a les propriétés suivantes :

Symétrie $\boldsymbol{X \perp Y|Z \Leftrightarrow Y \perp X|Z}$ (6)

Décomposition $X \perp Y \cup W|Z \Rightarrow X \perp Y|Z$ (7)

Union faible $X \perp (Y \cup W)|Z \Rightarrow X \perp W(Z \cup Y)$ (8)

Contraction $X \perp Y|Z \wedge X \perp W|(Z \cup Y) \Rightarrow X \perp (Y \cup W)|Z$ (9)

### 4.3.2. D-séparation

**Définition 3 :**

Soient (X,Y,Z) des nœuds du graphe G=(V,A). On dit que X et Y sont d-séparés par Z si pour tout chemin entre X et Y, l'une au moins des deux conditions suivantes est vérifiée :

- Le chemin converge en un nœud W, tel que $W \neq Z$, W n'est pas une cause directe de Z.
- Le chemin passe par Z, est soit divergent, soit en série au nœud Z.

X est d-séparé de Y par Z, et noté par $\langle X|Z|Y \rangle$.

**Définition 4 :**

Soient (X,Y) deux nœuds du graphe G=(V,A), et soit **Z** un ensemble de nœuds de ce graphe. Soit S un chemin entre les nœuds X et Y. On dit que S est d-séparé par **Z** si au moins l'une des deux conditions suivantes est vérifiée :

- Le chemin converge en un nœud w, tel que $W \notin \mathbf{Z}$ et $\forall Z \in \mathbf{Z}$, $W \notin C^{+}(Z)$ ou $C^{+}(Z)$ l'ensemble des ascendants de Z.

- Le chemin passe par un nœud $Z \in \mathbf{Z}$, et il est soit divergent, soit en série en ce nœud.

**Définition 5 :**

Soient (X,Y) deux nœuds du graphe G=(V,A), et soit **Z** un ensemble de nœuds de ce graphe. On dit que X et Y sont d-séparés par **Z** si tous les chemins entre X et Y sont d-séparés par **Z.**

**Définition 6 :**

Soient (**X**, **Y**, **Z**) trois ensembles des nœuds du graphe G=(V, A). on dit que **X** et **Y** sont d-séparés par **Z,** si tous les éléments de **X** sont d-séparés par **Z** de tous les éléments de **Y**.

**Remarque:**

*Deux variables X, Y sont d-séparées s'il existe une variable Z telle que tout chemin entre X, Y passe par Z.*

Dans le graphe de la figure (Fig.2.4), A et D sont d-séparés par B car le chemin orienté allant de A vers D passe par B. Et de même le nœud D d-sépare les nœuds ABC des nœuds EFG.

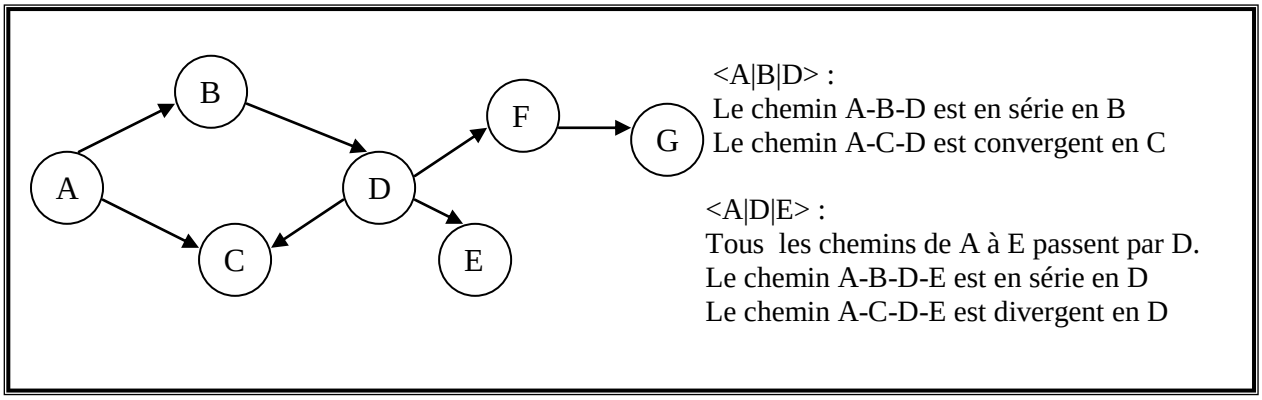



Fig.2.4. Exemple de D-séparation

## 5. Représentation probabiliste de la causalité

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 cette 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).

**Définition 7 :**

A et B sont indépendantes conditionnellement à C ssi :

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

-Le réseau bayésien permet de représenter graphiquement les indépendances conditionnelles.

Trois types de relations possibles entre A, B et C à savoir :

*Connexion en série :*

- A et B sont dépendants
- A et B sont indépendantes conditionnellement à C ssi :
  - P (C) est connue {P(A) n'intervient pas le calcul de P(B)}
  - P (S8|S5, S2) =P (S8|S4) =P (S8|Parents (S8)).

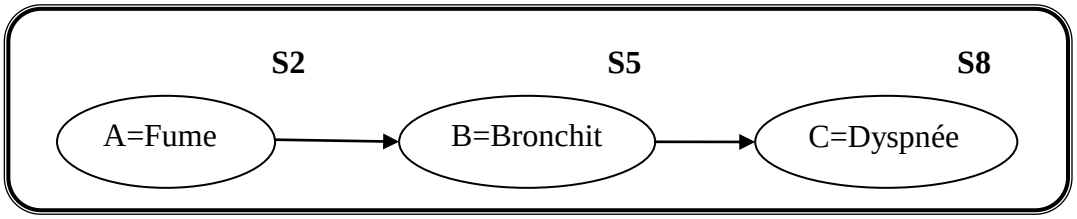



Fig.2.5. Connexion en série

*Connexion divergente :*

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

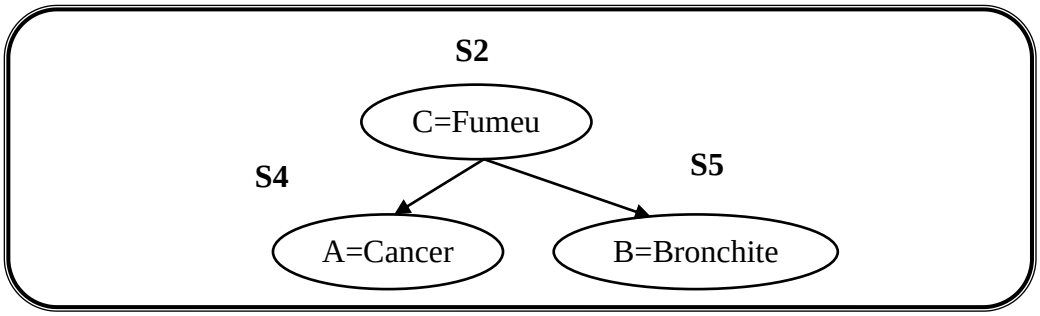



Fig.2.6. Connexion divergente

*Connexion convergente :*

- A et B sont indépendants.
- A et B sont indépendantes conditionnellement à C ssi :
  - ❖ P (C) est connue {P(A) n'intervient pas le calcul de P(B)}
  - ❖ P (S8|S6, S5) =P (S8|Parents (S8)).

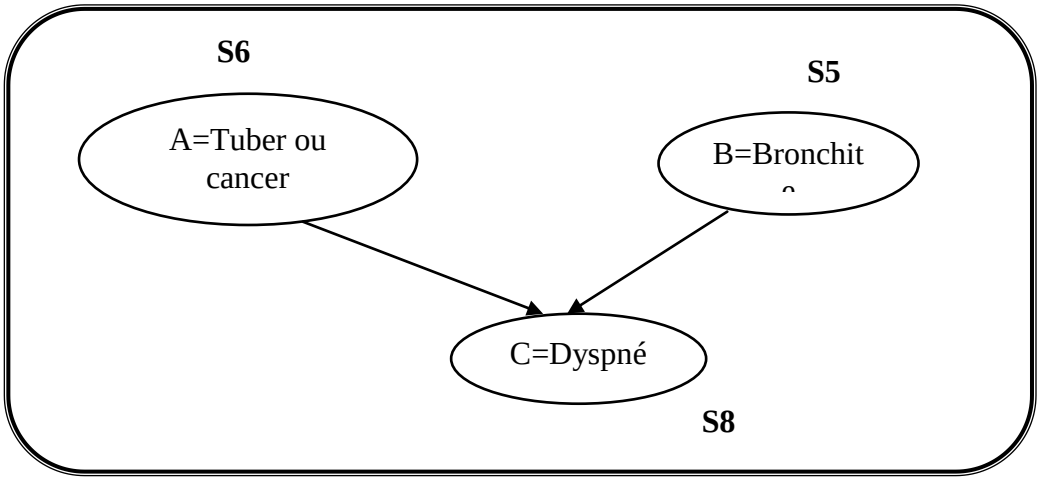



Fig.2.7. Connexion convergence

## 6. 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 y a 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 : étant 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] :

$$p(A|B)p(B) = p(A \cap B) = p(B|A)p(A) \qquad (10)$$

En notant $p(A \cap B)$ la probabilité que *A* et *B* aient tous les deux lieu. En divisant de part et d'autre par *P*(*B*), on obtient :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (11)$$

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 appelée 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*.

### 6.1. Autres écritures du théorème de Bayes

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

$$P(B) = P(A \cap B) + P(A^C \cap B) = P(B|A)P(A) + P(B|A^C)P(A^C) \qquad (12)$$

Afin de réécrire le théorème ainsi :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A^C)P(A^C)} \qquad (13)$$

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

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)} \qquad (14)$$

### 6.2. Exemple d'application de la formule de Bayes

L'application de la formule de Bayes est le cas le plus simple d'inférence bayésienne. Prenons l'exemple d'un test médical. Notons T la variable donnant le résultat du test (positif ou négatif) et M la variable représentant l'état de la personne (malade ou non malade). Le test est caractérisé par sa sensibilité ($Se$) et sa spécificité ($Sp$). La sensibilité nous donne la probabilité que le test soit positif sachant que la personne est malade [ROS 11]:

$$Se = P(T|M)$$

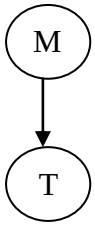



Fig 2.8. Représentation graphique de la relation entre la variable M (malade ou non malade) et la variable T (test positif ou négatif) dans le cas d'un test médical

La spécificité à l'inverse nous donne la probabilité que le test soit négatif lorsque la personne n'est pas malade :

$$Sp = P(\bar{T}\,|\bar{M}) \qquad (15)$$

La relation entre M et T est représentée graphiquement par la figure (Fig 2.8) :

La règle de Bayes nous permet de calculer la probabilité $P(M\,|T)$ que la personne soit malade sachant que le test est positif.

$$P(M\,|T) = P(T\,|M)\frac{P(M)}{P(T)} \qquad (16)$$

$$= P(T\,|M)\frac{p(M)}{P(T|M)P(M) + P(T|M)(1 - P(M))}$$

$$= \frac{Se\ p(M)}{Se\ P(M) + (1 - Sp)(1 - P(M)}$$

La probabilité $P(M)$ est la probabilité a priori que la personne soit malade. Elle est par exemple donnée par la prévalence de la maladie dans la population à laquelle appartient la personne.

Dans [MEY 09] les auteurs montrent l'importance et l'intérêt de la probabilité a priori dans le raisonnement bayésien au travers de l'exemple du diagnostic d'une sérologie positive pour le VIH, d'une part, chez une femme de 75 ans sans antécédents et, d'autre part, chez un toxicomane de 27 ans. Le test a les caractéristiques suivantes : Se = 0.97 et Sp = 0.98.

– Dans le cas d'une la femme âgée la prévalence de la maladie est P(M) = 1/500000, le calcul donne dans ce cas, P(M|T) = 0.00097.
– Dans le cas d'un toxicomane avec une prévalence P(M) = 1/10, nous obtenons P(M|T) = 0.982.

L'inférence bayésienne nous permet ici de fusionner l'information donnée par les caractéristiques du test avec celle donnée par la prévalence de la maladie. L'observation « test positif » est prise en compte pour calculer a posteriori, c'est à dire après observation du résultat du test, la probabilité que la personne soit malade. Cet exemple nous apprend que dans le cas de la femme âgée la spécificité du test n'est pas suffisante pour conclure à la présence de la maladie lorsque le test est positif.

## 7. Construction des réseaux bayésiens

La construction d'un réseau bayésien s'effectue en trois étapes essentielles, qui sont présentées sur la figure (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. Préconiser, dans un cadre général, l'une ou l'autre de ces approches serait pour le moins hasardeux [NAI 04].

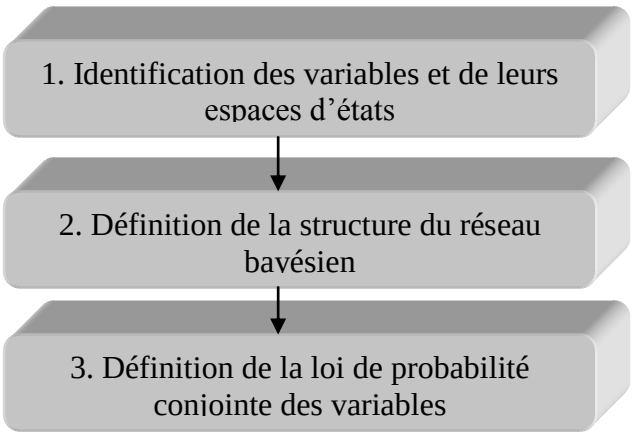



Fig.2.9. Etapes de construction d'un réseau bayésien

### 7.1. 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 $X_i$, 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 $X_i$, 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 [NAI 04].

### 7.2. Définition de la structure du réseau bayésien

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 $X_i$ influence-t-elle la variable $X_j$ ?. 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 $X_i$. L'expérience montre cependant que la représentation

graphique du réseau bayésien dans cette étape est un support de dialogue extrêmement précieux.

Un réseau bayésien ne doit pas comporter de circuit orienté ou boucle (Fig.2.10). Cependant, le nombre et la complexité des dépendances identifiées par les experts laissent parfois supposer que la modélisation par un graphe sans circuit est impossible. Il est alors important de garder à l'esprit que, quelles que soient les dépendances stochastiques entre des variables aléatoires discrètes, il existe toujours une représentation par réseau bayésien de leur loi conjointe. Ce résultat théorique est fondamental et montre bien la puissance de modélisation des réseaux bayésiens [NAI 04].

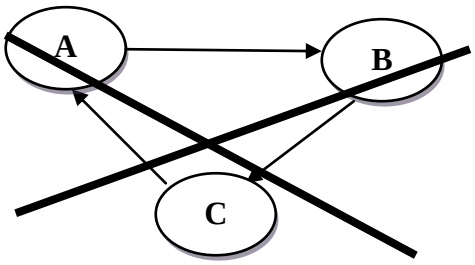



Fig.2.10. Boucle dans un réseau bayésien

Lorsque l'on dispose d'une quantité suffisante de données de retour d'expérience concernant les variables $X_i$, la structure du réseau bayésien peut également être apprise automatiquement par le réseau bayésien, à condition bien sûr que le logiciel utilisé soit doté de la fonctionnalité adéquate [NAI 04].

### 7.3. 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 $X_i$ dans le réseau bayésien :

- La variable $X_i$ n'a pas de variable parente: les experts doivent préciser la loi de probabilité marginale de $X_i$.
- La variable $X_i$ possède des variables parentes: les experts' doivent exprimer la dépendance de $X_i$ 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 $X_i$ peut être rencontré. La solution pragmatique consiste alors à affecter à $X_i$ 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 $X_i$.

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 $X_i$ peuvent être incorporées au modèle, afin d'affiner les probabilités introduites par les experts [NAI 04].

## 8. Inférence

L'inférence consiste à calculer la probabilité a postériori d'une (ou plusieurs) variable(s) du réseau conditionnelent aux variables observées : $P(X_i|O = o)$ où $O \subset X$ est l'ensemble des variables observées. L'inférence a été prouvée comme étant NP-difficile dans le cas général [COO 90][KOT 03]. L'inférence correspond au calcul de la probabilité d'une observation quelconque et il peut être exact ou approximative :

### 8.1. L'inférence exacte

Les techniques d'inférence reposent essentiellement sur l'indépendance des variables entre elles, ce qui permet de factoriser certaines parties du calcul et d'utiliser des techniques de programmation dynamique. Le cas des réseaux bayesien arborescents a été traité par Pearl [PEA 88] et permet par un simple passage de message entre les variables de calculer les différentes probabilités jointes. Le cas général a été traité par Jenson avec la construction d'un arbre de jonction [SHA 89]. Ces deux cas de réseau bayésien seront présentés dans les prochaines sections.

### 8.1.1. Propagation de messages (Algorithme Pearl)

Cette première méthode de propagation de messages [PEA 88] n'est valable que pour les réseaux bayésiens ayant comme graphe une arborescence. Cette particularité permet de déduire qu'il existe une unique chaîne entre deux sommets de ce graphe.

Soit O l'ensemble des variables observées. A partir d'un sommet X, nous pouvons définir deux ensembles de sommets $P_X$ et $E_X$ tels que $P_X$ regroupe les sommets de O dont la chaîne vers X passe par un parent de X et $E_X$ regroupe les sommets de O dont la chaîne vers X passe par un enfant de X et O = $P_X \cup E_X$ . Il est possible de montrer que :

$$P(X|O = o) \propto \lambda(X).\pi(X) \qquad (17)$$

Avec $\lambda(X) \propto P(E_X|X)$ et $\pi(X) \propto P(X|P_X)$.

Il s'agit ensuite de calculer des messages λ et π pour chaque variable selon un ordre topologique :

- **Les messages $\boldsymbol{\lambda}$**

Pour chaque enfant Y de X :

$$\lambda_Y(X = x) = \sum_{y \epsilon D_Y} P(Y = y|X = x).\lambda(Y = y) \qquad (18)$$

Le calcul de $\boldsymbol{\lambda}$ pour la variable X s'obtient à l'aide de la fonction suivante :

$$\lambda(X = x) = \begin{cases} \delta_{xO}^{x} & si\ X\ est\ instanciée\ à\ x_O\ (X \epsilon O) \\ 1 & si\ X\ est\ une\ feuille \\ \prod_{Y \epsilon\ enfants(X)} \lambda_Y(X = x)\ sinon. \end{cases}$$

Où le symbole de Kronecker $\delta_i^j$ a la signification suivante : $\begin{cases} 1\ si\ i = j \\ 0\ si\ i \neq j \end{cases}$

- **Les messages $\boldsymbol{\pi}$**

Pour l'unique parent Z de X :

$$\pi_X(Z = z) = \pi(Z = z) \prod_{U \epsilon\ enfants(Z) \setminus \{X\}} \lambda_U\ (Z = z) \qquad (19)$$

$$\pi(X=x)=\begin{cases}\delta_{xO}^{x} & si\ X\ est\ instanciée\ à\ x_O\ (X\epsilon O)\\ P(X) & si\ X\ est\ une\ racine\\ \sum_{z\epsilon D_Z} P(X=x|Z=z).\pi_X(Z=z) & sinon.\end{cases}$$

Dans ce cas sa complexité en temps est $O(nd^2)$ et en espace $O(nd)$ avec n=|X|, $d=max_{X_i\epsilon X}D_{X_i}$ cetet propagation peut également s'effectuer dans les polyarbres[4].

### 8.1.2. L'arbre de Jonction (Clique-tree propagation)

La méthode est applicable pour toute structure de graphe sans circuit contrairement à la méthode de propagation de messages précédente. L'idée est de fusionner les variables pour se ramener à un arbre. Elle se comporte de la façon suivante [LAU 88][JEN 90]:

*La phase de construction* : elle nécessite un ensemble de sous étapes permettant de transformer le graphe initial en 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 à l'inférence, mais qui reste équivalent au niveau de la distribution de probabilité représentée. Cette transformation se fait en trois étapes :

– La première étape est la **moralisation** du graphe G. Elle consiste à « marier »deux à deux les parents de chaque nœud, en les reliant par un arc non orienté. A l'issue de cette étape, il reste encore des arcs orientés entre chaque nœud et chacun de ses parents. On finit de moraliser le graphe en enlevant des directions de chaque arc orienté. On aboutit alors au graphe moralisé $G^m$.

– La deuxième étape est la **triangulation** du graphe $G^m$. Cette étape consiste à extraire de $G^m$ un ensemble de cliques de nœuds. Une clique est un sous-graphe du graphe $G^m$ dont tous les nœuds sont connectés deux à deux. Le graphe $G^t$ obtenu est triangulé quand l'ensemble de ses nœuds peuvent être éliminés. Un nœud peut être ´éliminé s'il appartient à une clique dans le graphe.

– Cette dernière étape correspond à la construction de l'arbre de jonction. A partir du graphe $G^t$ obtenu à l'issue de la triangulation, le problème est de calculer l'arbre couvrant de poids minimum. Pour ce faire, on va procéder à l'élimination des nœuds qui

[4] Graphe acyclique dont les sommets peuvent avoir plusieurs parents.

font partie d'une clique. Ce processus d''elimination n'est pas sans rappeler l'algorithme d'élimination de Bucket. L'arbre T obtenu est un arbre non orienté, dans lequel les nœuds sont des cliques.

*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 à l'ensemble du réseau, de manière à mettre à jour l'ensemble des distributions de probabilités du réseau. Ceci ce 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(U| ε) où U représente l'ensemble des variables du réseau bayésien et l'ensemble des nouvelles informations sur les dites variables. ε peut, par exemple, être vu comme un ensemble d'observations faîtes à partir de capteurs.
La construction de l'arbre de jonction est la dernière étape avant de procéder à l'inférence proprement dite. Nous rappelons que pour un réseau bayésien donné, l'arbre de jonction est construit une et une seule fois. Cette méthode a une complexité, en temps et en mémoire, exponentielle en la largeur d'arbre.

### 8.2. 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 variationnels qui permettent de briser les cycles du réseau. Pour faire de l'inférence, il suffit de résoudre un problème d'optimisation, i.e. minimiser la distance de Kullback-Leibler. D'une manière intuitive, cela revient à trouver la meilleure approximation de la distribution P(E), i.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 comme avec l'algorithme de Pearl.

## 9. 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 [NAI 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 [NAI 04].

- **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 [NAI 04].

- **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.

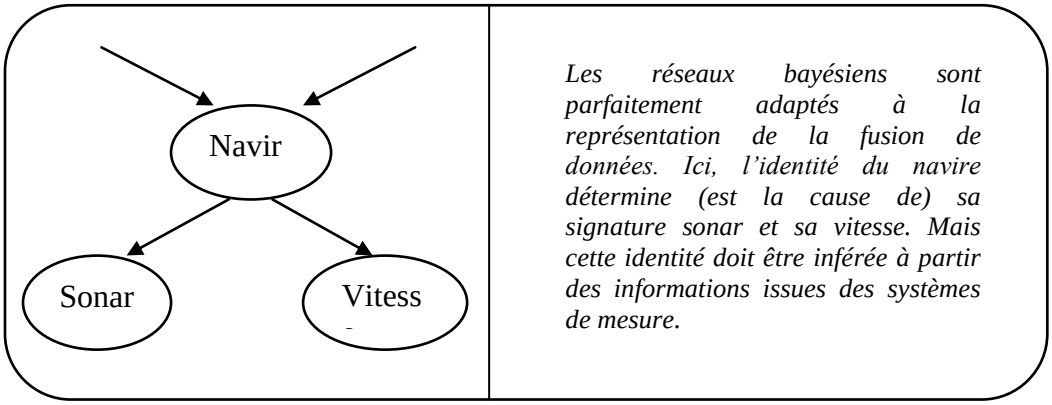



Fig.2.11. Principe de la fusion de données par réseau bayésien

Dans les années 1980 avec des systèmes experts, des applications comme l'analyse financière, le *scoring* ou la détection de fraudes ont été progressivement considérées comme relevant du domaine de la modélisation quantitative, et donc abordées par des techniques comme les réseaux neuronaux ou les arbres de décision, techniques quantitatives qui se révèlent incapables de prendre en compte par elles-mêmes la révision des modèles.

L'exemple de l'autorisation des transactions sur cartes bancaires est assez significatif. L'un des premiers systèmes experts développés dans ce domaine fut l'*Authorizer Assistant* d'*American Express*, au début des années 80 [NAI 04].

## 10. 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 [ABE 96]:

- 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.

### 10.1. 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 du 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].

### 10.2. 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 bayésien 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 bayésien 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 du décision et lui permet de gagne du temps puisqu'il existe moins de variables à interpréter [HEN 92].

## 11. Avantages des réseaux bayésiens

Le point clé dans l'utilisation de réseaux bayésiens (modèles graphiques) est dans le fait que l'analyse et la manipulation de modèles à plusieurs variables entraînant des relations de dépendance peuvent être considérablement facilitées en explorant les relations entre le modèle de probabilité et la représentation graphique. Les principaux avantages s'inscrivent dans la description du modèle et dans l'efficacité du calcul [SMY 97] [SOU 02]:

- **Description du modèle**

  Les graphes sont une façon naturelle de représenter l'information sous une forme compacte que l'humain peut saisir, comprendre et utiliser. En particulier, la structure d'un modèle graphique montre de manière simple et claire les dépendances conditionnelles dans le modèle de probabilité correspondant, permettant l'évaluation et la révision du modèle. De plus, le fait que le modèle graphique nécessite l'encodage explicite et la confrontation des hypothèses, peut être extrêmement utile dans la construction du modèle.

- **Efficacité du calcul**

  Les modèles graphiques sont une base solide pour la spécification d'algorithmes d'inférence efficaces pour effectuer des mesures requises dans le modèle de probabilité. L'avantage du calcul dans les modèles graphiques réside dans le fait que ces algorithmes d'inférence peuvent être spécifiés automatiquement dès lors que la structure initiale du graphe est déterminée. Il est à noter que le cadre des modèles graphiques ne permet pas automatiquement d'éviter l'explosion combinatoire des paramètres qui peut résulter de la construction de modèles plus réalistes. Il permet plutôt d'identifier une procédure d'inférence efficace de manière automatique si la structure du modèle le permet.

Du point de vue des applications, l'avantage des réseaux bayésiens est de pouvoir être utilisé aussi bien à partir d'une connaissance explicite que d'une connaissance implicite, ou de toute combinaison des deux, parmi leurs buts [ROS 11]:

- Intégrer la notion d'incertitude dans les systèmes experts, vu que la construction de celui ci nécessite presque toujours la prise en compte de *l'incertitude* dans le raisonnement.
- Leur capacité d'effectuer des inférences dans un contexte d'incertitude, en quelque sorte comme alternative aux systèmes experts.
- La combinaison des connaissances à priori et les connaissances empiriques (expérimentales), ce qui permettra l'enrichissement par l'expérience de la base de connaissances à priori par la connaissance empirique contenue dans les données.
- Quantifier la notion de « plus probable ».
- Nouveaux algorithmes **d'apprentissage**.
- Analyser d'autres algorithmes ne manipulant pas explicitement des probabilités.
- Ils permettent de rendre quantitatifs les raisonnements sur les causalités que l'on peut faire à l'intérieur du graphe.

Du point de vue des applications, les avantages et inconvénients des réseaux bayésiens par rapport à quelques-une des techniques concurrentes peuvent se résumer sur le tableau ci-dessous [NAI 04]. Nous avons regroupé avantages et inconvénients selon les trois rubriques suivantes (Tab.2.1) : l'acquisition, la représentation et l'utilisation des connaissances.

| Connaissances | Réseaux neurones | Arbres de décision | Systèmes experts | Réseaux bayésiens |
|---|---|---|---|---|
| **ACQUISITION** | | | | |
| Expertise seulement | | | * | |
| Données seulement | * | + | | + |
| Mixte | + | + | | * |
| Incrémental | + | | | * |
| Généralisation | * | + | | + |
| Données incomplètes | + | | | * |
| **REPRESENTATION** | | | | |
| Incertitude | | | + | * |
| Lisibilité | | + | + | * |
| Facilité | + | * | | |
| Homogénéité | | | | * |
| **UTILISATION** | | | | |
| Requêtes élaborées | | | + | * |
| Utilité économique | + | | | * |
| Performances | * | | | |

Tab.2.1. Avantages comparatifs des réseaux bayésiens [NAI 04]

La représentation adoptée est la suivante (Tab.2.1):

- A chaque ligne correspond une caractéristique, qui peut être un avantage, ou la prise en compte d'un problème spécifique.
- Si la technique considérée permet de prendre en compte ce problème, ou présente cet avantage, un signe + est placé dans la case correspondante.
- Un signe * est placé dans la case de la meilleure technique du point de vue de la caractéristique considérée.

## 12. Conclusion

Les réseaux bayésiens font partie de la famille des modèles graphiques. Ils regroupent au sein d'un même formalisme la théorie des graphes et celle des probabilités afin de fournir des outils efficaces autant qu'intuitifs pour représenter une distribution de probabilités jointe sur un ensemble de variables aléatoires. Ce formalisme très puissant permet une représentation intuitive de la connaissance sur un domaine d'application donné et facilite la mise en place de modèles performants et clairs.

La représentation de la connaissance se base sur la description, par des graphes, des relations de causalité existant entre des variables décrivant le domaine d'étude. A chaque variable est associée une distribution de probabilités locale quantifiant la relation causale.

Dans ce chapitre nous avons couvert les principaux aspects concernant les réseaux bayésiens. Nous avons vu que ces modèles graphiques sont des outils bien adaptés à la description des problèmes de décisions dans l'incertain, en permettant notamment de tenir compte des connaissances du problème à traité.

Le chapitre suivant est consacré à l'intégration des réseaux bayésiens dans l'approche de Raisonnement à Partir de Cas.

# Chapitre 3

# Intégration des Réseaux Bayésiens dans le Raisonnement à Partir de Cas

## Sommaire

# 1. Introduction

Le domaine des systèmes à base de connaissances au cours des années est devenu un domaine mature. Il est caractérisé par l'existence d'un ensemble de méthodes de représentation des connaissances, d'inférence et de raisonnement. De nombreuses applications ont été construites, qui sont utilisés au quotidien, et par conséquent ont prouvé la valeur des diverses méthodes pour les systèmes d'aide à la décision intelligente et d'autres applications. Et comme le domaine des méthodes individuelles devient plus exploré et mieux comprise, l'identification de limites et les points forts s'ouvre pour l'intégration des différentes méthodes de raisonnement dans les systèmes combinés.

Ce chapitre débute, à la section 2, par les différentes architectures CBR/BN et à la section 3, par un état de l'art des principaux systèmes intégrant les réseaux bayésiens dans le CBR. Cette étude a pour objectif l'analyse des propriétés de BN et CBR et comment peuvent-ils être combinés ?.

# 1. Les différentes architectures CBR/BN

Le Raisonnement à partir de cas (CBR) et les réseaux bayésiens (BN) peuvent être combinés dans les façons suivantes [BRU 09]:

- En parallèle
- Dans la séquence BN-CBR
- Dans la séquence CBR-BN

Dans l'architecture parallèle, les deux méthodes utilisent toutes les variables d'entrée et donc de produire une classification indépendante. Les résultats sont comparés par un algorithme "choisir le meilleur résultat", et la meilleure classification est choisie. Bruland [BRU 09] est mis sur des approches intégrées, représentées par les deux combinaisons séquentielles. BN et CBR sont reliés de telle sorte que le premier système dans la séquence calcule ce que le deuxième système a besoin. Les types de variables utilisés dans le domaine du problème sont les suivantes:

- $I_i$ est le numéro de variable d'entrée i. Pour les figures (Fig 3.1-4), les variables $I_1$ $I_2$ et $I_3$, sont uniquement utilisés par le BN et les variables $I_5$, $I_6$ sont uniquement utilisés par le système CBR. La variable d'entrée $I_4$ est utilisée par les deux systèmes.

- $A_j$ est le numéro de variable de médiation j. Les variables médiatrices représentent des concepts dans le modèle de domaine. Un expert du domaine peut également être une partie du processus de classification et il peut établir l'évidence d'une variable médiatrice.
- *D* est une variable qui est dérivée par inférence à partir de la connaissance du domaine. Elle est le principal résultat de BN dans la séquence BN-CBR. Elle peut être la solution d'un cas, comme un résultat intermédiaire dans l'architecture de séquence CBR-BN.
- *C* est une variable de classification et elle peut être calculé par un BN ou être la solution finale d'un cas.

L'utilisateur crée une description de la requête du problème avec les variables d'entrée. Deux spécialisations de chaque type de séquence ont été développées. L'architecture BN-CBR-1 est représentée dans la figure (Fig.3.1). Les identificateurs de cas sont présentés sous forme de variables du réseau bayésien et ils ont les valeurs binaires on/off qui indique si le cas est activé ou non. La variable dérivée dans l'ensemble de variable D est à l'origine des cas d'être activé. Ces Ds dérivées qui sont obtenues à partir des variables d'entrée par inférence sur la base de connaissances du domaine.

Par conséquent, le BN a un rôle de filtrage dans la phase de remémoration du système CBR. Les mesures de similarité sont uniquement appliquées sur les cas filtrés. Les systèmes sont faiblement couplés dans cette architecture, parce que l'information utilisée dans le BN est caché dans le système CBR. L'architecture BN-CBR-1 a été trouvée dans deux travaux de recherche [AAM 98] [GOM 04].

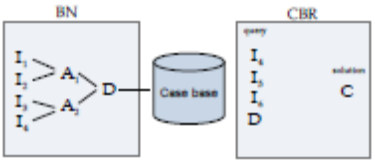



Fig.3.1. Architecture BN-CBR1

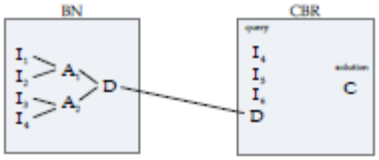



Fig.3.2. Architecture BN-CBR2

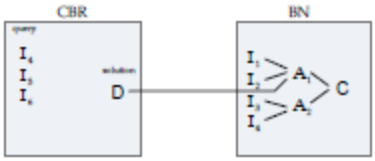



Fig.3.3. Architecture CBR-BN1


Fig.3.4. Architecture CBR-BN2

L'architecture BN-CBR-2 est illustrée dans la figure (Fig.3.2) et ici, les systèmes sont étroitement couplés. L'utilisateur dispose de toutes les variables d'entrée et les variables [1..4] sont définis comme éléments d'évidence dans le BN, et l'expert mis son évidence au BN. Les probabilités du réseau sont calculées et le D est placé dans la description du cas avec I1, I2, I3, I4. La phase remémoration dans CBR est effectuée, aboutissant à une liste classée des solutions de cas similaires (variable de classification). Une variante de l'architecture BN-CBR-2 a été trouvée parmi les travaux de recherche [AAM 98]. Dans ce travail, le système CBR est le maître et le BN est l'esclave. La connaissance de domaine est représenté par le réseau bayésien et un réseau sémantique où les variables du BN sont partagés avec le réseau sémantique. Dans cette approche, le système CBR utilise le réseau bayésien en plusieurs étapes du processus de raisonnement. Par exemple, l'étape d'activation dans remémoration (les étapes sont du moteur explication [AAM 94]) définit certaines évidences au BN qui active les cas pertinents (BN-CBR-1 architecture). L'étape expliqué en remémoration trouve les cas les plus similaires. L'étape accent dans remémoration établit de nouvelles évidences du cas au BN et trouve de nouvelles probabilités occasionnelles qui peuvent renforcer l'information dans le réseau sémantique. Le BN peut aussi être utilisé dans la phase de Réutilisation.

L'architecture CBR-BN-1 est représentée dans la figure (Fig.3.3) et elle montre deux systèmes étroitement couplés. Le système CBR trouve une liste des n-meilleures avec les variables d'entrée $I_4$, $I_5$, $I_6$, et la solution de cas contient la variable dérivée D. La variable D, les variables d'entrée I1, I2, I3, I4 sont définis comme éléments d'évidence dans le BN, avant que les probabilités a posteriori de la variable de classification sont calculées. Il y a deux façons d'envisager l'architecture CBR-BN-1. La première est celle où le système CBR est une étape de prétraitement pour le BN. La deuxième est celle où le BN est utilisé dans la phase de réutilisation de CBR. Dans la première approche, l'étape de prétraitement peut être utilisée sur une partie du modèle BN qui est inconnu. Ici, un expert peut créer des cas qui remplacent ce modèle BN inconnu. Les cas doivent contenir les variables D avec des valeurs probabilistes. Certains peuvent également être donnés par un expert du domaine qui est présent dans le processus de classification. Une fois les variables sont insérées à titre d'évidence dans le BN, la variable de classification est calculée. Si les valeurs de C sont à portée les unes des autres il y a une possibilité de plus d'une classe probable. Si la meilleure valeur C est inférieur à un seuil n'y a pas de classe probable. Dans le second cas de l'architecture CBR-BN-1, la phase Réutilisation peut contenir un raisonnement sous incertitude. Le système CBR trouve les cas les plus similaires et la phase Réutilisation peut utiliser le BN afin d'adapter le cas. La

variable de classification est disponible dans le BN. L'architecture CBR-BN-1 a été trouvée dans un des travaux de recherche [TRA 08].

L'architecture CBR-BN-2 est illustrée dans la figure (Fig.3.4). Le système CBR utilise les variables d'entrée $I_4$, $I_5$, $I_6$ pour trouver une solution qui contient le modèle BN le plus approprié. Le modèle BN est chargé et les variables d'entrée $I_1$, $I_2$, $I_3$, $I_4$ sont fixés comme évidence. Par la suite, la variable de classification est calculée. Les modèles différents BN disposent d'une évidence commune et des nœuds de classification, mais d'autres nœuds, les liens de causalité, et les tables de probabilités conditionnelles peuvent être différentes dans chaque modèle. Les informations utilisées dans le système CBR sont cachées du système BN, et par conséquent les systèmes sont faiblement couplés dans cette architecture. L'architecture CBR-BN-2 a été trouvée dans un des travaux de recherche [PAV 09].

## 2. Etat de l'art des principaux systèmes intégrants le BN dans le CBR

Il y a plusieurs façons d'intégrer le réseau bayésien (BN) dans l'approche de Raisonnement à Partir de Cas (CBR). Cependant, une remémoration et adaptation bayésienne dans un système CBR sont les contributions principales de ce travail. Ci-dessous nous avons identifié quelques travaux de recherches, qui sont présentées avec une description brève comment ils combinent le BN et le CBR.

### 3.1. Le système Creek

Le premier système Creek, dans lequel la connaissance de domaine générale a été représentée par un réseau sémantique [AAM 94], a été étendu avec un composant bayésien et donné un exemple en trouvant la cause du problème "d'une voiture ne commence pas"[AAM 98]. Le réseau sémantique contient des liens causals avec l'incertitude et le raisonnement probabiliste est exécuté par un réseau bayésien (Fig.3.5). Les nœuds et les relations causales sont partagés entre le réseau sémantique et le BN. Les cas sont présentés comme des variables (marche/arrêt) dans le BN. Le système Creek utilise le BN pour choisir des cas pertinents pour appliquer la mesure de similarité sur la remémoration de cas bayésien. Le BN est une étape de prétraitement intervient dans la phase de remémoration. Le réseau bayésien peut aussi calculer les relations causales qui sont utilisées dans la phase de réutilisation.

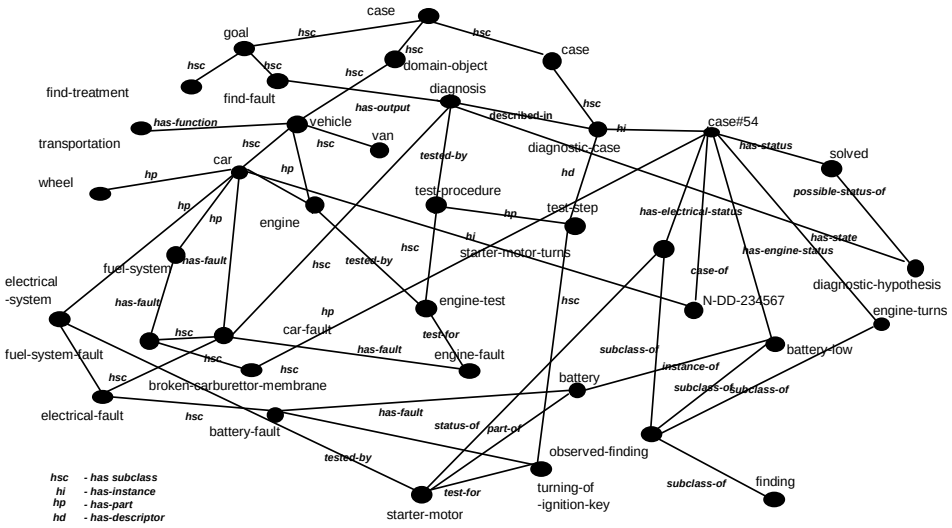



Fig.3.5. Un réseau sémantique de concepts Creek [AAM 94]

## 3.2. Le système INBANCA

Aha D et al [AHA 96] décrivent une architecture nommée INBANCA (**In**tegrating **B**ayesian **N**etwork with **CA**sed-Based Reasoning) dans lequel des réseaux bayésiens et le raisonnement à partir de cas coopèrent sur les tâches planifiant multiagent. La figure (Fig.3.6) résume la structure de contrôle pour le système INBANCA. Dans bref, INBANCA utilise des réseaux bayésiens pour choisir des actions et le CBR pour choisir un cas représentant les instruments de plan cette action. Les étapes du plan sont exécutées dans l'état actuel jusqu'à ce que le but de l'action réussisse (par exemple en fonctionnant à l'achèvement sans échec) ou échoue. Cette boucle sélection-exécution continue jusqu'à ce qu'un certain critère arrêtant n'est satisfait (par exemple un but est marqué ou un temps mort). L'étude a lieu après que chaque étape action est exécutée. Il implique la mise à jour des réseaux en réponse à la sélection d'action. La mise à jour de cas basée sur le retour d'information pour appliquer leur action planifie et mettre à jour la base de cas en modifiant la fonction de récupération de cas par exemple. En ajoutant des nouveaux cas ou modifiant les indices existants de nouveaux cas sont produits comme des expériences accumulées.

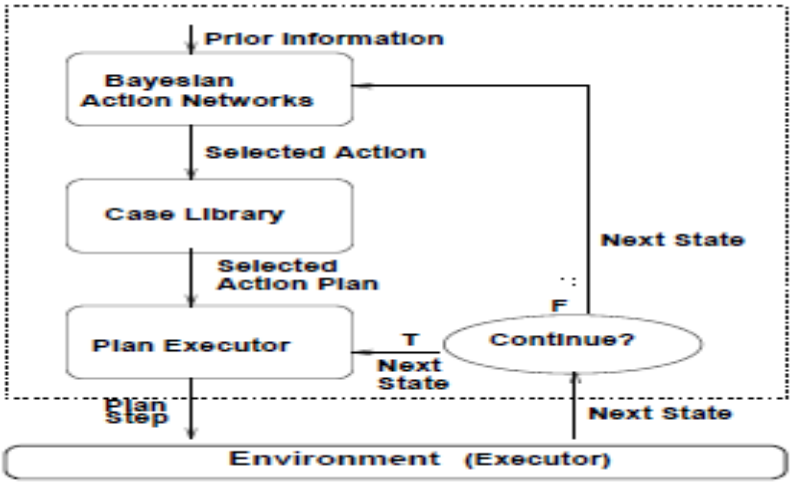



Fig.3.6. Architecture de control INBANCA [AHA 96]

### 3.3. Le système développé par Silvia et al.

Silvia et al [SIL 00] présentent un exemple d'utilisation des réseaux bayésiens en combinaison avec la technique de raisonnement à partir de cas (Fig.3.7). Chaque requête de l'utilisateur est prise comme un cas. Chaque nœud du réseau représente un attribut ou une caractéristique utilisée dans une requête. Les arcs décrivent les relations entre les attributs comme étant les fréquences d'apparitions des attributs dans les requêtes de l'utilisateur. Ces fréquences représentent également l'importance des attributs pour le client. Lorsqu'une nouvelle requête arrive, elle est stockée sous forme de cas et pour chaque attribut un nœud est ajouté dans le réseau. Ensuite, les liens entre ces caractéristiques sont rajoutés et les fréquences d'apparition sont mise à jour.

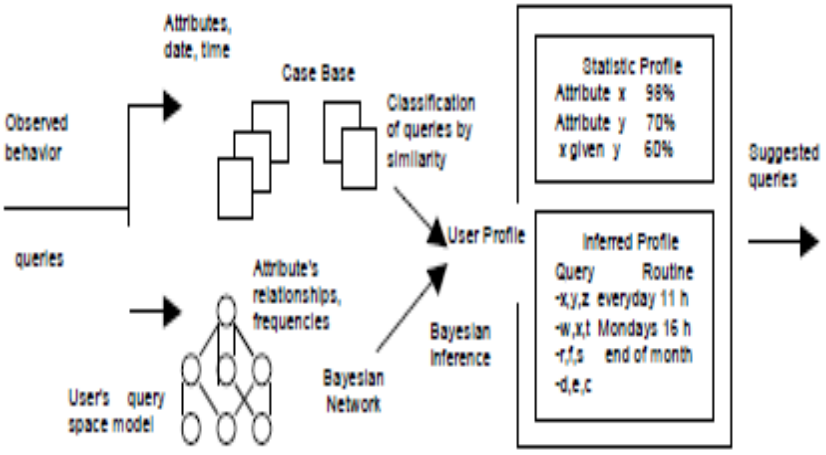



Fig.3.7. Intégration BN et CBR dans le profil utilisateur [SIL 00]

### 3.4. Le système Bayesian Case Reconstruction (BCR)

La Reconstruction de Cas Bayésien (BCR) [HEN 02] est un système de conception d'expériences de dépistage de cristallisation macromoléculaire. La connaissance du domaine est incomplète, la base de cas est incomplète, il y a un grand nombre de variables ayant un grand nombre de valeurs, et il ya limitation de temps, de matériel et de ressources humaines. BCR est utilisé pour étendre la couverture de la bibliothèque de base de cas. Le BN est construit à partir des experts du domaine et le contenu de la bibliothèque de cas. La phase remémoration sélectionne les cas les plus probables, et ils sont désassemblés afin de former de nouvelles solutions. Le réseau bayésien contient les relations de cause à effet à partir du modèle de domaine qui sont bien comprises. Dans la phase de réutilisation, le BN est utilisé pour trouver les nouvelles solutions les plus probables. Le résultat est une solution plausible, seulement.

### 3.5. Le système développé par Gomes P

Gomes P [GOM 04] présente un outil de génie logiciel qui aide des ingénieurs logiciel à réutiliser des designs précédents. Le système combine CBR, BN et WordNet. Les cas dans le système ont une partie de description de problème contenant un certain nombre de l'ensemble de synonyme WordNet. Les cas sont des nœuds dans le réseau bayésien, comme sont aussi l'ensemble de synonyme de la description de problème. L'ensemble de synonyme de la description de problème sont utilisés pour trouver tous les parents de la relation d'hyperonyme de WordNet (est - un) et tous les parents sont insérés dans le BN. Les tables de probabilité conditionnelles sont construites avec des formules selon le nombre de nœuds parents. La phase de remémoration est exécutée dans trois étapes comme suit : a) la description de cas de question est utilisé pour activer l'ensemble de synonyme dans le réseau bayésien, b) les nœuds du BN sont calculés et les cas les plus pertinents sont trouvés et c) leurs probabilités sont utilisées pour classer les cas.

### 3.6. Le système développé par Tran H et al

Tran H et al [TRA 08] décrivent un système CBR distribué utilisée pour trouver des solutions dans le domaine défaut de communication du système. La description du problème est un ensemble de symptômes, *S*, et la solution du problème contient une hypothèse de défaut *H*. Leur processus de raisonnement comporte deux étapes: classement et sélection. L'étape de classement (phase de remémoration) trouve les cas les plus similaires à leurs relations du BN

$S_i|H_j$. L'étape de sélection (phase réutilisation), utiliser les relations BN $S_i|H_j$ des cas pour construire un réseau bayésien. L'hypothèse la plus probable à partir de BN est choisie.

### 3.7. Le système développé par Pavon F et al

Un autre système qui combine BN et CBR est utilisé pour choisir les paramètres optimaux pour un algorithme utilisé dans des domaines différents [PAV 09]. Le BN est appris et évalué à partir des expériences dans les domaines avec les algorithmes utilisant des fixations de paramètres différents. La phase de remémoration choisit les cas les plus similaires. La phase d'adaptation est utilisée pour calculer une mesure de précision dans le complément pour calculer les arguments les plus probables du BN. Les cas les plus fiables sont ceux qui ont un grand nombre d'expériences et un grand nombre de variations dans les paramètres utilisés.

### 3.8. Le système développé par Dong D et al

Dong D et al examinent le raisonnement à partir de cas probabiliste en intégrant des réseaux bayésiens avec CBR. Ils proposent un cadre CBR probabiliste pour la gestion de prescription d'obésité (PCOPM) pour aider des professionnels de santé à partager leurs expériences de prescription d'exercice d'obésité en ligne (Fig.3.8). Le PCOPM lie ensemble CBR et BN dans un cadre unifié qui inclut tant expérience d'obésité que l'incarnation intelligente de prise de décisions pour la gestion d'obésité. Cette approche facilite la recherche et le développement de gestion d'obésité basée sur le web intelligent [DON 10].

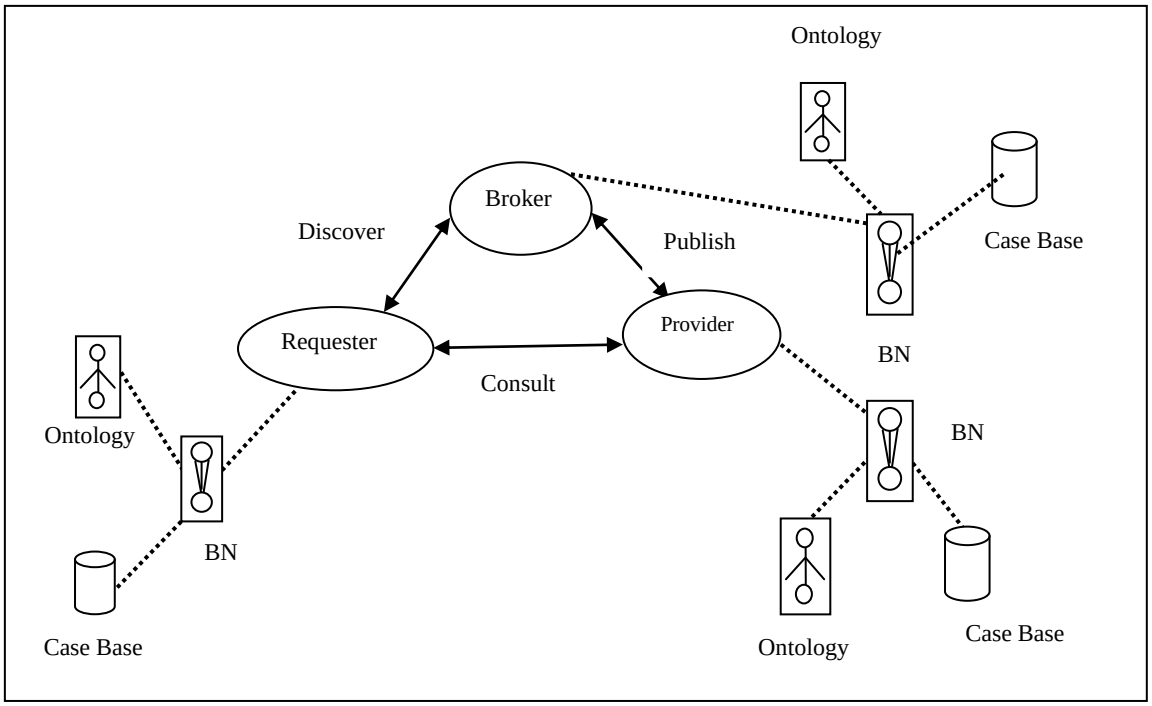



Fig.3.8. PCOPM : une architecture CBR probabiliste pour le management de prescription d'obéisité [DON 10]

### 3.9. Le système développé par Gravem A

[GRA 10] présente un système d'aide à la décision qui combine le Raisonnement à Partir de Cas et les réseaux bayésiens. Ce système a été conçu pour l'utilisation dans le soin palliatif, mais a été développé avec des données alimentaires et de vin à cause des retards de la préparation de bases de données médicales. Il est concentré sur le développement d'une intégration de CBR et BN qui peut utiliser les points forts de ces méthodologies (Fig.3.9). Ce système utilise un réseau bayésien pour représenter des relations générales entre la nourriture et des raisins. De plus, il a une base de cas indépendante qui représente des recettes. En utilisant le réseau bayésien pour réduire l'espace de recherche pour le module CBR. Ce système a pu trouver des solutions nouvelles et intéressantes. Cette conception utilise les avantages des deux méthodologies de raisonnement et pouvoir ainsi surmonter les limitations d'une seule approche. Le problème le plus apparent avec ce système est le réseau bayésien. Ce réseau a été revendiqué pour être trop général. Un expert de vin devrait donc être révisé. Ce système a été testé et évalué par un expert de vin dans la tâche de trouver un vin adéquat pour un repas donné et avec une précision de 87 % sur une base de 12 cas.

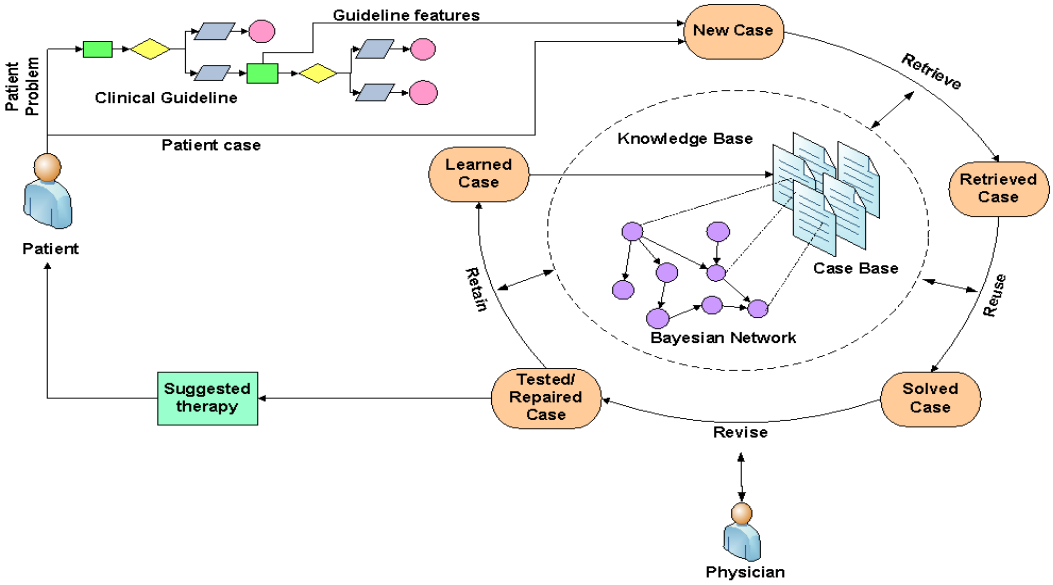



Fig.3.9. Architecture du système proposé par Gravem [GRA 10]

## 3.10. Le système ABM

Ocampo M et al [OCA 11] se concentrent sur la construction d'outils de support de diagnostic pour la Méningite Bactérienne Aiguë ABM, rapportant une évaluation comparative de la qualité d'un Système d'aide à la Décision Clinique (CDSS) résultant de l'application de Raisonnement à Partir de Cas, à celui d'un système CDSS utilisant un système expert bayésien (Fig.3.10). Bien que les deux approches se soient avérées être utiles, celui basée sur des techniques CBR montre quelques capacités intéressantes comme la précision plus haute, l'étude automatique ou la conquête d'expérience et aussi une meilleure réponse au manque de données en entrée. Les trois systèmes développés fonctionnent avec les hauts niveaux d'exactitude proposent par exemple le diagnostic correct basé sur un certain ensemble de symptômes mais celui basé sur le CBR présente quelques capacités supplémentaires qui semblent très prometteuses pour mettre en œuvre cette sorte de systèmes dans un scénario du monde réel.

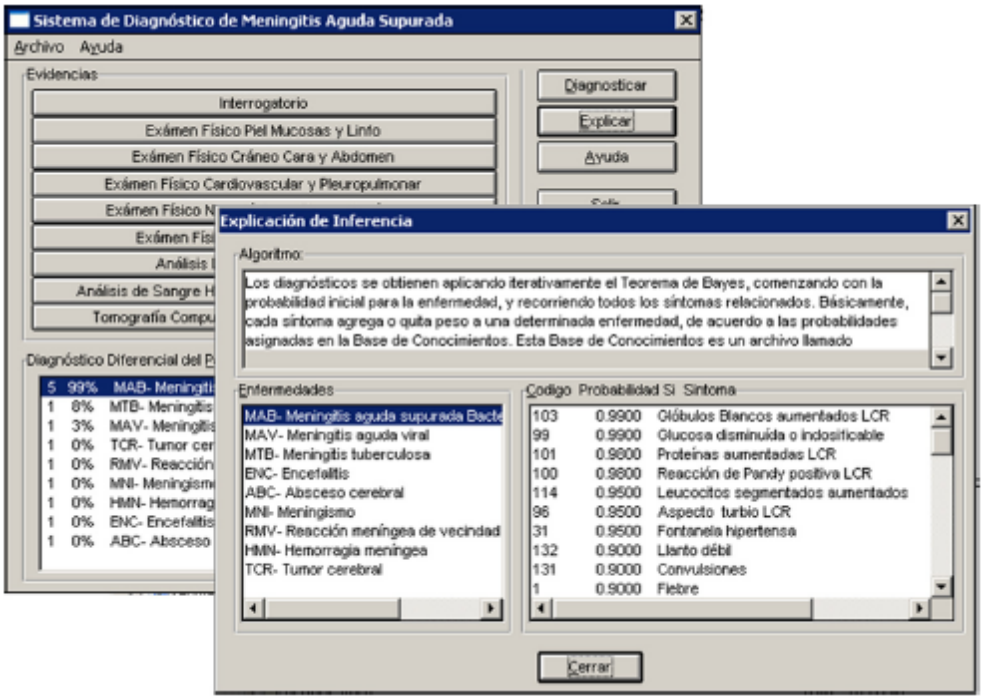



Fig.3.10. La fenêtre principale du système AMBDE [OCA 11]

## 4. Discussion

Les systèmes hybrides décrit dans la section 3, se différent par rapport à notre approche proposée essentiellement dans la façon d'implémentation et le domaine d'application. Seuls les systèmes proposés dans [GRA 10], [DON 10] et [OCA11] traitent le domaine médical. Le système de [GRA 10] a utilisé le BN pour représenter des dépendances entre des variables d'états et il a inclus aussi des directives cliniques. Ce système veut tester un réseau qui représente la connaissance experte généralisée, tandis que la base de cas est composée de cas patients, peut renforcer le raisonnement et produire de bons résultats. Le système décrit dans [DON 10] combine le BN avec le CBR pour la gestion de prescription d'obésité et pour aider des professionnels de santé à partager leurs expériences de prescription d'exercice d'obésité en ligne. Le système de [OCA 11] dédie au diagnostic pour la Méningite Bactérienne Aiguë ABM. Trois autres systèmes qui ont intégré une troisième méthode avec la combinaison CBR/BN, par exemple dans le système Creek [AAM 94] les connaissances sont partagées entre un réseau sémantique et un réseau bayésien. Le réseau bayésien est intervient dans la phase d'adaptation. Le système INBANCA de [AHA 96] utilise les réseaux bayésiens et le CBR pour la planification des tâches du multiagent tel que le BN sert à choisir des actions et le CBR pour choisir les instruments du cas. Le système proposé par [GOM 04] a utilisé le WordNet avec la combinaison CBR/BN pour construire un outil d'aide au génie

logiciel qui aide les ingénieurs de réutiliser des logiciels de design précédents. Concernant l'architecture CBR/BN, des six systèmes combinant CBR/BN, parmi eux, quatre ont utilisé le CBR et le BN en parallèle [SIL 00] [PAV 09] [GRA 10] [OCA 11]. Deux ont utilisé le CBR et le BN en série [HEN 02] [TRA 08]. D'autres systèmes ont utilisé un troisième module séquentiellement avec la combinaison CBR/BN comme par exemple le réseau sémantique [AAM 94], le multiagent [AHA 96], le WorldNet [GOM 04] et l'ontologie [DON 10].

Notre approche est basée sur l'utilisation parallèle de CBR et BN en parallèle afin d'améliorer le processus décisionnel en connaissances incertaines. Il n'y a pas un grand volume de publication qui décrit une combinaison de BN avec l'approche CBR dans le domaine médical. La différence que dans notre approche, le BN représente la base de cas et il décrit les dépendances entre les descripteurs de cas (partie problème et partie solution). La plupart des autres systèmes que nous avons étudiés ont vraiment aussi utilisé la base de cas pour créer le réseau bayésien comme [HEN 02] [TRA 08]. Aucune de ces approches n'est similaire à notre approche. Nous voulons utiliser le BN comme une mémoire de cas en l'utilisant pour rechercher le cas remémoré le plus facile à adapter.

## 5. Notre démarche

Notre démarche dépend de l'organisation de la mémoire choisi. De ce fait, nous avons adopté l'organisation hiérarchique, et comme le domaine de notre travail est la médecine où l'incertitude est un compagnon permanent de l'activité médicale, nous avons basée sur la modélisation de la base de cas par un réseau bayésien.

Le cas est décrit en termes de dimension attribut. Ces attributs sont des données cliniques, des données biologiques, des caractéristiques extraites à partir de l'image du foie [DJE 04], et de diagnostic conclu avec sa thérapie. Par définition un cas est une entité qui contient diverses informations représentant un domaine particulier. Un cas contient les données empiriques décrivant l'expérience acquise dans la résolution d'une situation précise.

Nous aborderons au chapitre 4 (section 5.2) la description du cas de notre approche dédié au diagnostic des pathologies hépatiques. A travers cette description, nous avons modélisé la mémoire de cas par un réseau bayésien, puis nous décrivons la phase de la remémoration et d'adaptation. La figure ci-dessous (Fig.3.11) montre le schéma général du système d'aide au diagnostic proposé. Le système se compose de trois parties principales : une base de cas, une phase de remémoration et une phase d'adaptation. A l'entrée du système, le cas cible est représenté par un ensemble de paramètres décrivant des symptômes se

manifestant chez un patient et décrivant une image échographique et scannographique du foie. Des informations pouvant aider le médecin dans le processus de prise de décision, représentent la sortie de ce système. Ces informations, qui peuvent être des cas similaires au cas cible donné, des diagnostics potentiels, etc., seront synthétisées et présentées au médecin sous la forme d'un diagnostic avec sa thérapie.

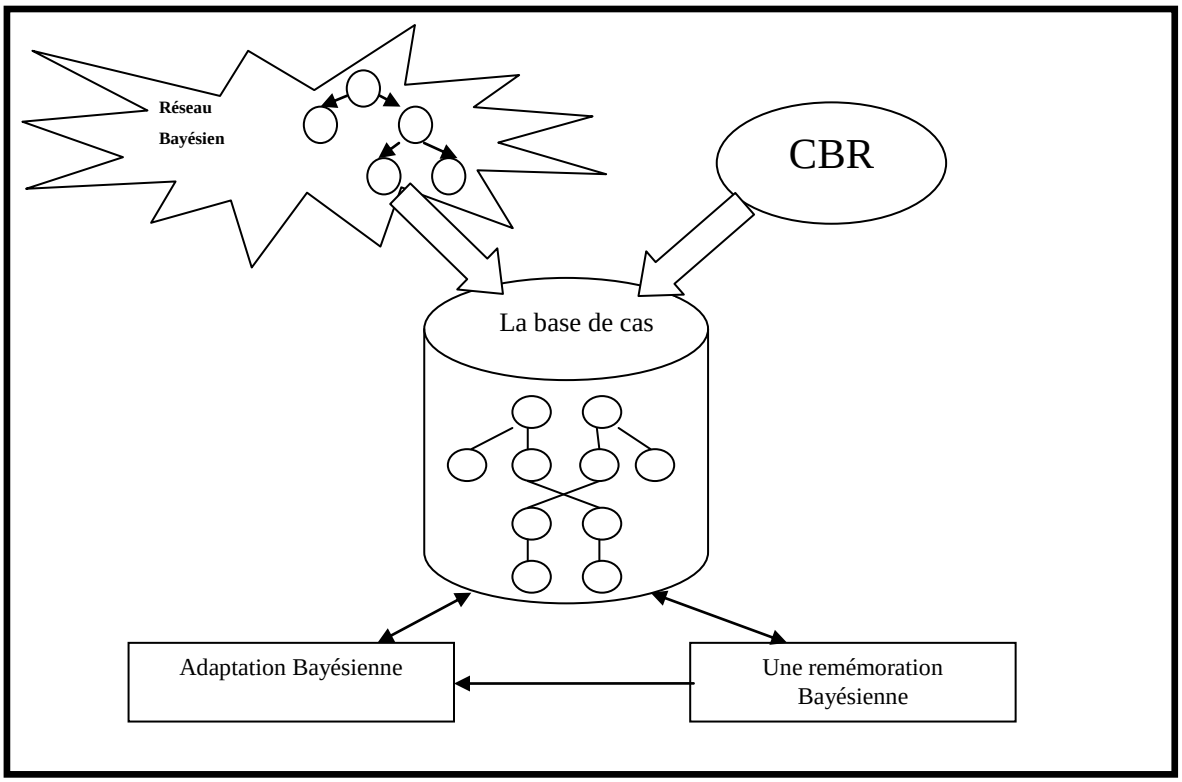



Fig.3.11. Schéma générale de notre approche

## 6. Conclusion

L'objectif de notre travail est de proposer un système d'aide au diagnostic des pathologies hépatiques. Afin de se situer par rapport aux travaux existants et de choisir les techniques appropriées pour notre système, nous avons étudié dans ce chapitre les différentes architectures intégrants les réseaux bayésiens dans le RàPC et un état de l'art concernant les principaux systèmes combinant le CBR/BN.

Le RàPC a été choisi comme méthode de résolution de problèmes dans notre système d'aide au diagnostic des pathologies hépatiques et les réseaux bayésiens comme un outil de modélisation des connaissances du RàPC. Nous avons déduit de cette étude plusieurs conclusions importantes pour la suite de notre travail. Nous avons confirmé notre démarche

de modélisation de la mémoire de cas par un réseau bayésien présentée dans ce chapitre par une mise en parallèle de l'intégration de réseau bayésien dans le RàPC. A travers cette intégration, nous détaillons dans le chapitre suivant les fonctionnalités de la phase de remémoration et la phase d'adaptation.

# Chapitre 4

# Modélisation bayésienne de la remémoration et de l'adaptation pour l'aide au diagnostic médical

## Sommaire

# 1. Introduction

Le diagnostic est l'opération qui consiste à trouver les causes d'un phénomène connaissant un ensemble d'observations, c'est-à-dire l'opération qui consiste à répondre à la question : pourquoi a-t-on obtenu le résultat que l'on vient d'observer ? Ceci est l'acception la plus générale de la notion de diagnostic. Dans le cas du diagnostic médical, diagnostiquer revient à trouver les causes qui ont engendrées un certain nombres de symptômes chez le patient. Les causes sont dans ce cas la maladie ou l'affection dont souffre le patient [LEB 94][AGN 98].

Les systèmes de diagnostic doivent gérer des connaissances incertaines, et raisonner en environnement incertain en utilisant des mesures particulières de l'incertitude. Les réseaux bayésiens représentent un modèle particulièrement efficace, car ils permettent une représentation au sein d'un même modèle de connaissances qualitatives causales et de connaissances quantitatives exprimant l'incertitude que l'on a sur les connaissances qualitatives. En effet, de nombreuses applications sont basées sur l'utilisation d'un réseau bayésien en diagnostic médical et essentiellement sur le diagnostic à partir d'observations incertaines. Les modèles utilisés sont soit construits à partir d'une expertise humaine, soit construits par différents algorithmes qui exhibent les relations causales contenues de façon intrinsèque dans les données.

Dans le cadre de ce travail, le réseau bayésien et notamment les probabilités sont données par l'expert du domaine. Nous avons choisi ce modèle pour modéliser la base de cas afin de mieux cerner les différentes facettes de la mise en œuvre d'un système RàPC et son modèle bayésien de mémoire. Nous présentons dans ce chapitre les deux premières grandes étapes du RàPC : la remémoration et l'adaptation. Nous détaillons comment, et avec quelles mesures, la recherche de similarité a été effectuée. En suite, nous présentons la phase d'adaptation ainsi qu'une mesure d'adaptation pour adapter la solution du cas remémoré afin que le cas cible étudié trouve aussi une solution et soit ainsi résolu.

# 2. Description générale du problème médical

Le problème médical pris en compte dans ce travail est le diagnostic des pathologies hépatiques. Le raisonnement médical part d'un examen clinique et d'un ensemble de données interrogatoires. En effet, à partir des symptômes observés exemple douleurs, alcool, hypertrophie, ictère,…, le médecin identifie les facteurs prédisposant. Sur la base de ces

facteurs plusieurs hypothèses sont établies et seront confirmés par un examen biologique. Soulignons qu'aucun examen clinique n'est pas spécifique à 100%. Il reste toujours une incertitude, et ce n'est qu'à travers l'examen radiologique que le médecin arrive à poser son diagnostic final. La figure (Fig.4.1) schématise les différentes étapes de diagnostic.

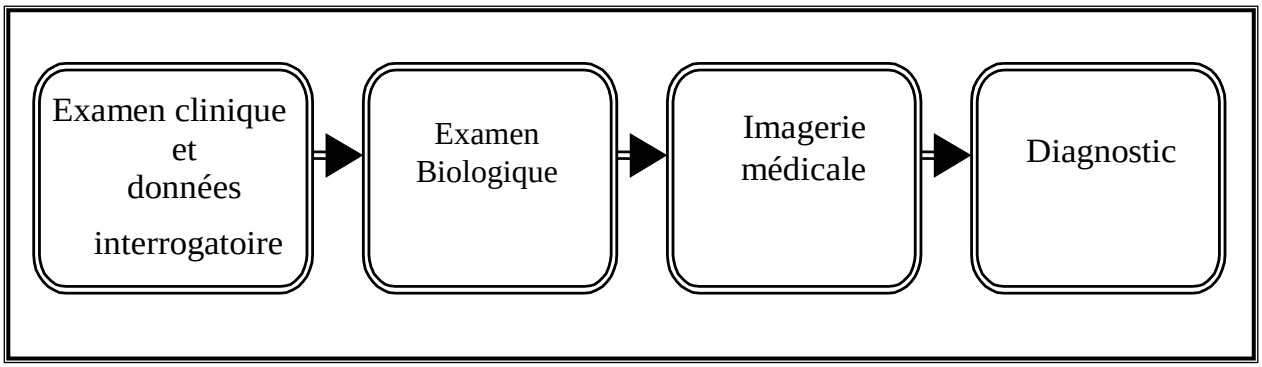



Fig.4.1. Description du problème du diagnostic

## 3. Base de cas

La base de cas est une collection de cas de résolution du même problème. Elle tient une place primordiale dans la mise en place d'un RàPC de qualité. Le problème de la représentation des cas commence toujours par le choix d'informations que l'on veut stoker, trouver la structure correspondante, organisation des cas et l'indexation [KOL 93][AAM 94].

Dans notre problème de diagnostic, les observations sur lesquelles se basent la décision médicale sont imparfaites, de grande qualité et incertaines dans le processus décisionnel, qu'il soit diagnostic, thérapeutique ou pronostique, c'est donc un processus sous incertitude. C'est pour cette raison, que pour diagnostiquer des pathologies hépatiques, il nous parait nécessaire de modéliser les connaissances par un réseau bayésien.

## 4. Réseaux bayésiens et diagnostic médical

### 4.1. Introduction

Un diagnostic médical est le résultat du raisonnement d'un médecin, décision très souvent prise à partir d'informations incertaines et/ou incomplètes. En amont de ce raisonnement, il faut aussi être capable de modéliser ces informations incertaines.

Une méthode consiste à se placer dans le cadre de la théorie des probabilités, ce qui nous amène tout naturellement aux réseaux bayésiens proposées par Pearl [PEA 88] dans les

années 80, retrouvés parfois sous le nom de systèmes experts probabilistes. Les réseaux bayésiens possèdent de nombreux avantages (modélisation probabilistes de l'incertitude, possibilité de raisonnement dans le sens symptômes-diagnostic) qui font d'eux des outils privilégiés dans le cadre du diagnostic médical. Ces avantages ont été présentés au chapitre 2 (section 11).

Un réseau bayésien est un modèle représentant des connaissances incertaines sur un phénomène complexe et permettant à partir des données un véritable raisonnement. Il permet également de représenter un ensemble de variables aléatoires pour lesquelles on connaît un certain nombre de relations de dépendances [BEL 02][DEL 05].

Dans notre problème, les observations sur lesquelles se base le diagnostic des pathologies hépatiques sont très imparfaites. Elles peuvent être incertaines car les connaissances cliniques sont l'expression d'observations statistiques sur des échantillons de patients présentant des maladies plus au moins fréquentes, ayant des formes cliniques différentes et ne s'exprimant pas toujours par la même symptomatologie, partageant certains signes avec d'autres maladies ou présentant des réponses variables à un traitement donné. Le processus décisionnel de diagnostic est donc un processus sous incertitude.

Les réseaux bayésiens sont des outils efficaces de représentation des connaissances incertaines et ils décrivent hiérarchiquement toutes les étapes du raisonnement du médecin.

### 4.2. Définition des variables du réseau

Pour utiliser le formalisme des réseaux bayésiens, nous avons exhibé les informations utiles à la représentation de notre problème. Suite aux discussions faites avec les experts du domaine, nous avons déduit que les informations les plus pertinentes à traiter sont réparties en quatre niveaux à savoir (Fig.4.2):

- Niveau clinique : Il est constitué d'un ensemble d'attributs (facteurs) possibles permettant la détection des différentes maladies hépatiques, exemple : Alcool, Douleur, Ictère, Hépatomégalie, …
- Niveau biologique : Il définit la phase biologique. Il présente tout les états d'analyse médicale pouvant être détectés à partir du niveau précédant. Exemple : Bilirubine, Phosphatase, Sérologie, Hyperglobuline, …

- Niveau imagerie médicale : Il comprend les caractéristiques de l'image échographique et scanographique du foie par exemple : Taille, Homogénéité, Densité, périmètre, nombre de lésion, l'angle aigu [DJE 04], …
- Niveau diagnostic et thérapie : il représente le diagnostic final avec la thérapie : Kyste hépatique, Abcès, Angiome, Métastases hépatiques,…

Ces informations sont représentées par des variables à valeurs discrètes. A chaque variable est associée une probabilité spécifique et une table représentant l'estimation des probabilités conditionnelles. Les variables du dernier niveau diagnostic et thérapie représentent les maladies du foie et qui sont réparties en classes, à savoir:

- Classe tumeur hépatique maligne (Tumeur M) : Carcinome hépatocellulaire, Carcinome fibrolamellaire, Hépatoblastome, Angiosarcome, Hépatome, Adénome, Lipome,…
- Classe Tumeurs hépatique bénignes (Tumeur B) : Hémangiomes, Hyperplasie nodulaire focale, Carcinome Hépathocellulaire, Localisation hépatique des hemopathie,…
- Classe Cancers secondaires (Métastases secondaires hépatiques) : Carcinomes, cancers endocrines, cardia, les métastases secondaires suite à mélanomes, les mélanomes choroïdiens,…
- Classe Kyste : Kyste hépatique, Amylase hépatique, foie polykystique, …
- Classe Abcès : Abcès hépatique, …
- Autre classe : Cirrhose, foie cardiaque, Stéatose hépatique, Glycogenome, ….

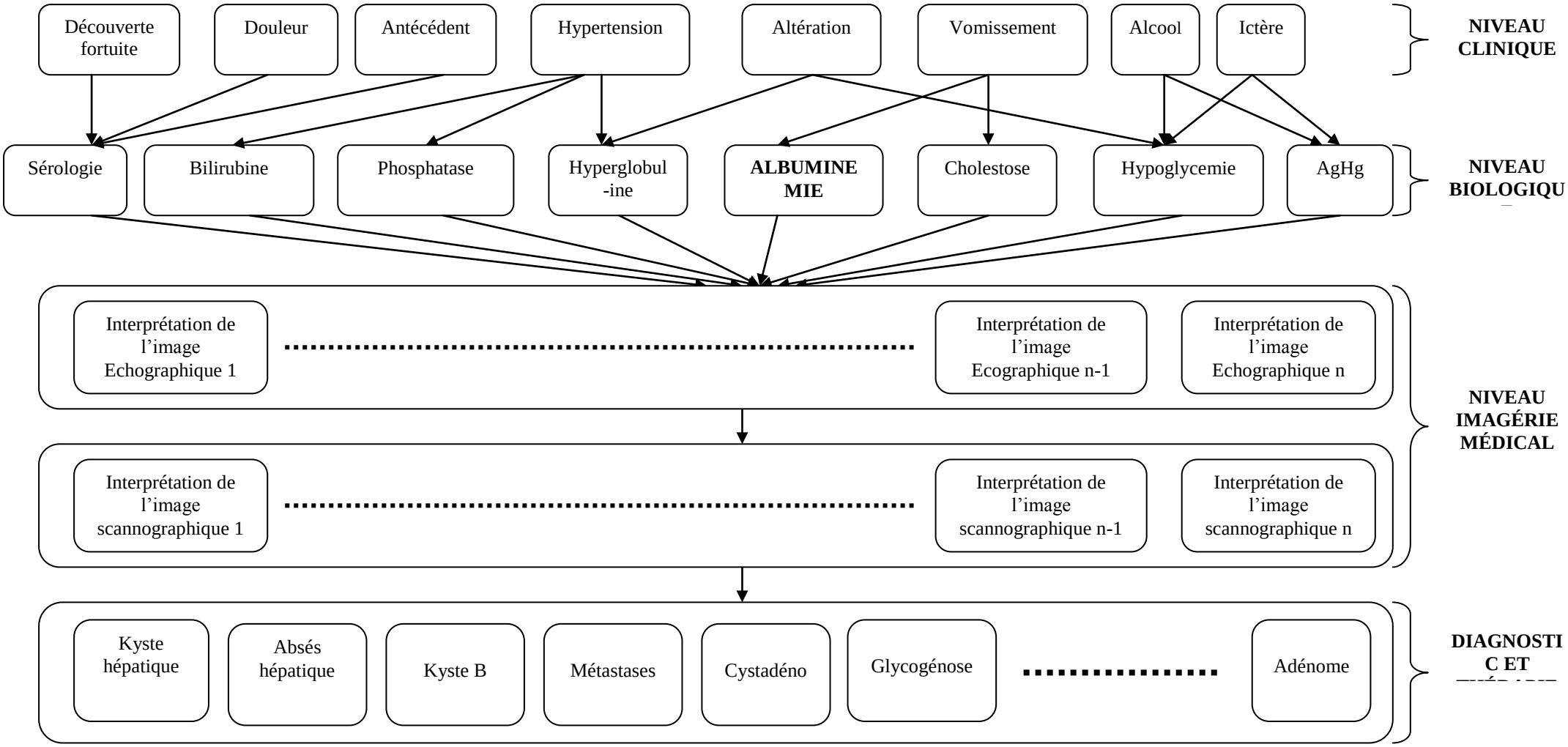



Fig.4.2. Schéma général d'une modélisation de la base de cas par un BN pour l'aide au diagnostic des pathologies hépatiques

### 4.3. Exemple de réseau bayésien

La maladie du foie peut être causée par une ou plusieurs causes. On voudrait que la base de cas puisse résoudre des situations différentes. Le médecin voudrait pouvoir déterminer les probabilités d'existence des différentes maladies connues à connaître après avoir donné les valeurs observées qui ne sont autre que les symptômes.

Pour compléter l'information sur le réseau bayésien, les lois de probabilité sont données à chacune de ses variables aléatoires. Ces lois, dans ce cas, des variables discrètes, sont des tables de probabilités.

La construction de ce réseau (Fig.4.3) garantit une série d'indépendances conditionnelles vérifiant localement sa sémantique. Elle est issue de l'algorithme [PAC 04].

---

**Algorithme 1** : Algorithme de construction du réseau bayésien.

---

Choisir un ordre sur les variables $X_1, \ldots, X_n$

**Pour** i=1 à n **Faire**

- Ajouter $X_i$ au réseau
- Sélectionner ses parents dans $X_1, \ldots, X_{i-1}$ tels que

  $P(X_i \mid Parents(X_i)) = P(X_i \mid X_1, \ldots, X_{i-1})$

**Fin Pour**

---

**Remarque**

*Il est préférable d'avoir un modèle causal, c'est-à-dire qu'il est meilleur d'ajouter la cause « racine » en premier et ensuite les variables qui sont influencées par la cause* [PAC 04].

Dans la figure (Fig.4.3), nous donnons l'exemple d'un réseau bayésien qui modélise de manière raisonnable le processus de diagnostic de deux maladies du foie. Le mode de raisonnement est ici représenté par un diagramme de causalité (BN). Les probabilités sont déterminées par l'expert du domaine [DJE 06].

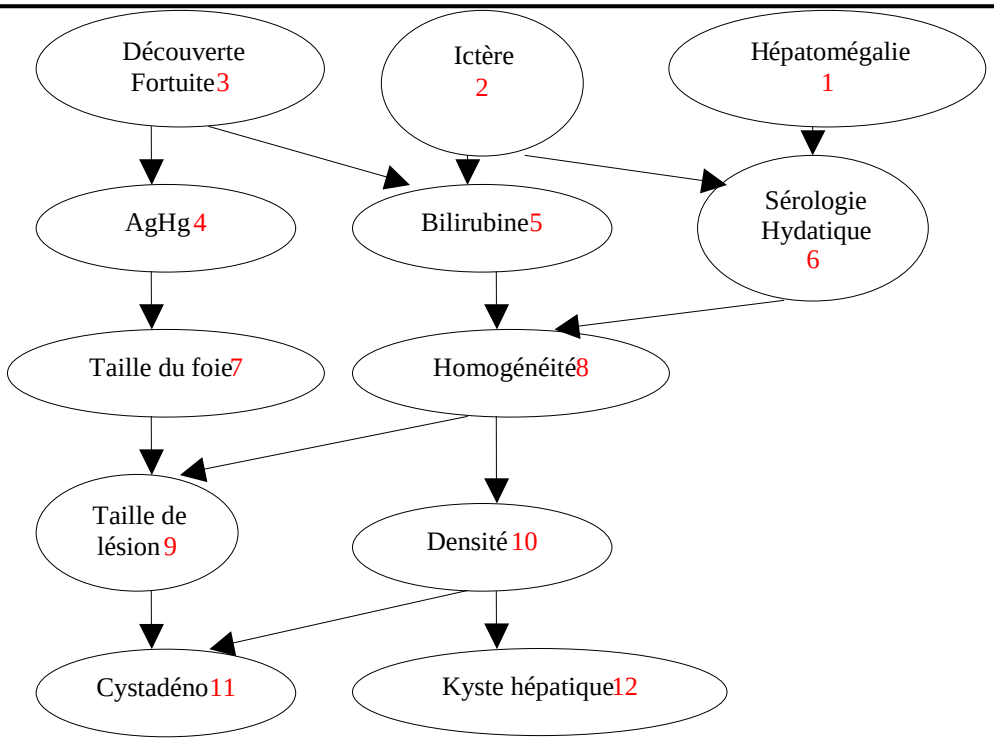



Fig.4.3. Exemple d'un réseau bayésien avec 2 maladies du foie

## 4.4. Inférence

L'efficacité de l'inférence dans les réseaux bayésiens dépend de plusieurs données : l'algorithme choisi et ses paramètres (en particulier le nombre d'échantillons), le réseau bayésien et ses caractéristiques (taille du réseau, « densité » du réseau, taille des domaines de définition des variables, etc.), et le calcul d'inférence à exécuter et ses caractéristiques (en particulier le nombre et la position des variables observation et cibles).

### 4.4.1. Exemple de propagation d'un message par l'algorithme Pearl

Pendant l'exécution de l'algorithme Pearl [PEA 88], chaque noeud X passe par cinq états différents :

1. *Attente de messages* : en notant $N_x$ le nombre de ses voisins, tant que X a reçu moins de ($N_x$ – 1) messages, rien ne produit.
2. *Calcul de messages de collecte* : X a reçu ($N_x$ – 1) messages, il est donc capable de calculer le message vers le seul voisin Y qui ne lui a rien envoyé. D'une manière générale, on dira que X est en phase de collecte.
3. *Attente de réponse*: X est en attente d'un message de son dernier voisin.
4. *Calcul de messages de distributions* : X a reçu le dernier message. Il est en mesure de calculer λ(x), p(x) et P(x|e). Il est aussi en mesure de distribuer les ($N_x$ - ) messages

qu'il n'a pas encore envoyés.

5. *Fin* : X est au repos. L'algorithme est terminé.

Notons que :

$N_X$ : l'ensemble des nœuds parents observés de X.

$D_X$ : l'ensemble des nœuds enfants observés.

La figure (Fig.4.4) montre un nœud X, l'ensemble de ses nœuds parents $N_X$ et l'ensemble de ses nœuds enfants $D_X$.

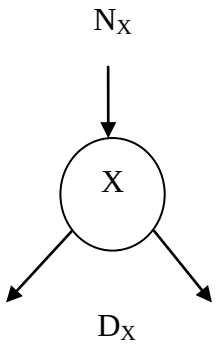



Fig.4.4. Sous-ensemble d'un arbre : un nœud, son parent et ses enfants

**Exemple**

Dans ce paragraphe, nous allons appliquer l'algorithme de Pearl sur un exemple d'un réseau bayésien de diagnostic des maladies du foie représenté par la figure (Fig.4.5) :

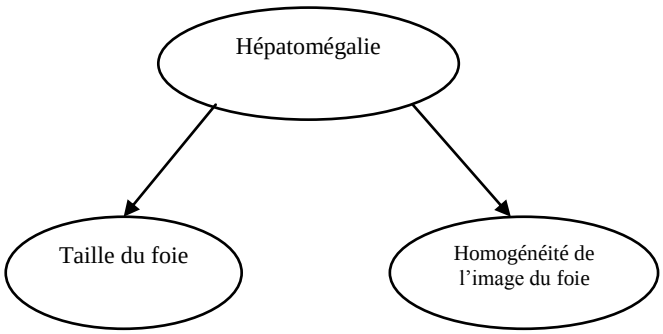



Fig.4.5. Exemple d'un réseau bayésien à 3 variables

La description des nœuds est donnée ci-dessous :

- Le nœud *Hépatomégalie* représente la variable aléatoire discrète *H*. Cette variable prend ses valeurs dans le domaine {*petit=*$h_0$*, moyen=*$h_1$*, grand=*$h_2$}.
- Le nœud *Taille du foie* représente la variable aléatoire discrète *TF*. Cette variable prend ses valeurs dans le domaine {*Normal=O, Anormal=N*}.

- Le nœud *Homogénéité de l'image du foie* représente la variable aléatoire discrète *HF*. Cette variable prend ses valeurs dans le domaine {*Homogène=O, Non homogène=N*}.

Les tables de probabilités conditionnelles des variables *H*, *TF* et *HF* sont données dans les tableaux Tab.4.1, Tab.4.2 et Tab.4.3 respectivement.

**Arrivée d'une nouvelle observation**

On suppose qu'on a une nouvelle observation de la valeur $h_2$ sur la variable *H*. par contre, on n'observe pas du tout les autres valeurs $h_0$ et $h_1$.

Cette observation s'appelle *évidence* et on écrit $e_H$={0.0.1}.

On souhaiterait savoir comment les autres variables vont réagir étant donnée cette observation.

| *H* | *P(H)* |
|---|---|
| $h_0$ | 0.30 |
| $h_1$ | 0.60 |
| $h_2$ | 0.10 |

Tab.4.1. Table de probabilité de la variable *H*

| *P(HF\|H)* | *O* | *N* |
|---|---|---|
| $h_0$ | *0.85* | 0.15 |
| $h_1$ | *0.50* | 0.50 |
| $h_2$ | *0.05* | 0.95 |

Tab.4.2. Table des probabilités conditionnelles de la variable *HF* étant donnée la variable *H*

Pour ce faire, on va propager le message que fournit l'évidence. Commençons par calculer les messages au niveau de la variable observée, car le calcul des messages $\boldsymbol{\lambda}$ et $\pi$, au niveau des variables observées, est un cas trivial :

**Calcul des messages $\boldsymbol{\lambda}$ et $\pi$ au nœud observé *H* :**

- Calcul des messages $\boldsymbol{\lambda}$ et $\pi$ :

  *H* est observé. On a donc $\boldsymbol{\lambda}(H)=\pi(H)=\begin{bmatrix}0\\0\\1\end{bmatrix}$

- Le message à transmettre à l'enfant HF est calculé comme suit :

$$\pi_{HF}(H) = \pi(H).\lambda_{TF}(H) = \begin{bmatrix}0\\0\\1\end{bmatrix}.\lambda_{TF}(H) = h_2\,)$$

Ou, $\lambda_{TF}(H = h_2) = \sum_{TF} p(TF = f|H = h_2).\lambda(TF = f) =$

$p(TF = O|H = h_2).\lambda(TF = O) + p(TF = N|H = h_2).\lambda(TF = N) =$

$0.90 * 1 + 0.10 * 1 = 1$

Donc $\pi_{HF}(H) = \begin{bmatrix}0\\0\\1\end{bmatrix}.1 = \begin{bmatrix}0\\0\\1\end{bmatrix}$

- De même, le message à transmettre à l'enfant *TF* est calculé comme suit :

$$\pi_{TF} = \pi(H).\lambda_{HF}\,(H) = \begin{bmatrix}0\\0\\1\end{bmatrix}$$

On peut ensuite calculer les messages au niveau des deux autres nœuds *HF* et *TF*. Les deux autres nœuds sont des feuilles de l'arbre. Le calcul des messages $\boldsymbol{\lambda}$ sera donc trivial au niveau de ces nœuds. On peut traiter *HF* et *TF* dans n'importe quel ordre car aucun de ces nœuds n'est racine de l'arbre, et aucun de ces nœuds n'est observé.

| *P(TF\|H)* | *O* | *N* |
|---|---|---|
| $h_0$ | *0.20* | 0.80 |
| $h_1$ | *0.75* | 0.25 |
| $h_2$ | *0.90* | 0.10 |

Tab.4.3. Table des probabilités conditionnelles de la variable *TF* étant donnée la variable *H*

**Calcul des messages $\boldsymbol{\lambda}$ et $\pi$ au nœud observé *HF*:**

- Calcul du message $\boldsymbol{\lambda}$ :

*HF* est une feuille donc on a $\lambda(HF) = \begin{bmatrix}1\\1\end{bmatrix}$

- Calcul du message $\pi$ :

*HF* n'est ni une variable observée, ni la racine de l'arbre. On a donc :

$\pi(HF) = p(HF|H).\pi_{HF}(H) = \begin{bmatrix}0.85 & 0.5 & 0.05\\0.15 & 0.5 & 0.95\end{bmatrix}.\begin{bmatrix}0\\0\\1\end{bmatrix}$, soit $\pi(HF) = \begin{bmatrix}0.05\\0.95\end{bmatrix}$

Enfin, on a $p(HF|H = e_H\,) \propto \lambda(HF).\pi(HF) = \begin{bmatrix}0.05\\0.95\end{bmatrix}$

On peut conclure que l'observation d'un foie grand (l'hépatomégalie= grande) peut

conduire à une image de foie non homogène.

**Calcul des messages $\lambda$ et $\pi$ au nœud observé *TF*:**

- Calcul du message $\boldsymbol{\lambda}$ :

*TF* est une feuille donc on a $\lambda(TF) = \begin{bmatrix}1\\1\end{bmatrix}$

- Calcul du message $\pi$ :

*TF* n'est ni une variable observée, ni la racine de l'arbre. On a donc :

$\pi(TF) = p(TF|H).\pi_{TF}(H) = \begin{bmatrix}0.20 & 0.75 & 0.90\\0.80 & 0.25 & 0.10\end{bmatrix}.\begin{bmatrix}0\\0\\1\end{bmatrix}$, soit $\pi(HF) = \begin{bmatrix}0.90\\0.10\end{bmatrix}$

Enfin, on a $p(TF|H = e_H) \propto \lambda(TF).\pi(TF) = \begin{bmatrix}0.90\\0.10\end{bmatrix}$

On peut conclure que l'observation d'un foie grand (l'hépatomégalie= grande) peut conduire à une taille grande de l'image de foie.

### 4.4.2. Modèles log-linéaires

Les modèles log-linéaires peuvent aussi être utilisés pour simplifier le nombre de paramètres d'une loi de probabilité conditionnelle, ou plus généralement la loi de probabilité jointe $P(Y,X_1,X_2, ...,X_n)$ d'une variable et de ses parents. Le principe, très général, de ces modèles est de décomposer le logarithme d'une loi de probabilité en une somme de termes décrivant les interactions entre les variables. Cette décomposition est dite saturée lorsque tous les termes sont présents dans la décomposition, et non saturée lorsque des hypothèses supplémentaires sont ajoutées, comme par exemple le fait que certaines variables soient indépendantes, pour supprimer des termes dans la décomposition [KEY 02][BLU 06][DEC 07]. Dans le cas qui nous intéresse, nous savons aussi que les parents sont mutuellement indépendants. De plus, nous proposons (comme [COR 03]) de ne garder que les termes d'interaction d'ordre inférieur ou égal à 2 ($u$, $u_i$, $u_{0i}$), arrivant au modèle log linéaire non saturé suivant :

$$\log P(Y, X_1, X_2, \dots, X_n) = u + \sum_i u_i(X_i) + \sum_i u'_i(X_i, Y) \qquad (1)$$

La détermination de ces termes d'interaction passe par la résolution d'un système linéaire, en utilisant certaines contraintes comme le fait que la somme des $P(Y,X_1, ...,X_n)$ doit être égale à 1. En supposant que l'expert soit interrogé sur toutes les probabilités marginales $P(x_i)$, $P(y)$, et sur toutes les probabilités conditionnelles $P(y \mid x_i)$ et $P(y|x_i)$, [COR 03] montre qu'il reste encore $2^n - 2n$ contraintes à satisfaire pour déterminer complètement les

paramètres du modèle log-linéaire. Nous considérons le modèle log linéaire saturé. On note $X_1$, $X_2$,….,$X_m$, les variables du modèle. Ces variables sont discrètes. La probabilité jointe, tenant compte de toutes les interactions possibles s'écrit comme suite [QUI 93][CHR 97]:
Notre réseau bayésien, établit précédemment, a au maximum n=2 parents. Pour ce modèle comportant un nœud, noté B avec n parents, notés $A_1$, ..,$A_n$, il faut définir $2^n$ probabilités conditionnelles et n probabilités a priori. Si n >=5, la tâche de l'expert devient complexe. L'idée est donc de se ramener à un modèle log linéaire simple. Cette approche permet d'obtenir une modélisation plus générale, mais nécessite davantage d'estimations de la part de l'expert lorsque le nombre de parents d'une variable est important.

## 5. Intégration des réseaux bayésiens dans le CBR

### 5.1. Modélisation de la base de cas par un réseau bayésien

Dans ce travail, la base de cas est modélisée par un réseau bayésien dont les paramètres du graphe se résument aux probabilités marginales et conditionnelles. La structure et les probabilités servant à l'inférence du modèle proviennent des avis d'experts. Ceci est justifié par le fait que les réseaux bayésiens sont généralement utilisés comme des systèmes experts capable de capitaliser et de modéliser la connaissance. Puis, au fur et à mesure que le temps passe et que l'expérience grandit, des données de retour d'expérience sont répertoriées. Ces données doivent obligatoirement enrichir le réseau bayésien. Ainsi, les avis d'experts qui ont servi à évaluer les probabilités du graphe, sont dans cette optique considérées comme les probabilités a priori dans un cadre probabiliste bayésien classique. Avec les données de retour d'expérience, la probabilité a posteriori est calculée, via la formule de Bayes (voir chapitre 2, section 6).

La construction de notre base de cas se décompose en trois étapes distinctes :

- ✓ Etape qualitative : elle ne tient compte que les relations d'influence existantes entre les variables;
- ✓ Etape probabiliste : dans laquelle, on doit introduire l'idée de distribution jointe sur les variables et on fait correspondre la forme de cette distribution au graphe crée;
- ✓ Etape quantitative : elle consiste simplement à spécifier numériquement les distributions de probabilités conditionnelles.

Notre travail porte sur la construction d'un réseau bayésien pour modéliser la base de cas. Ce réseau constitué d'un ensemble d'informations pour le diagnostic. Les nœuds du

réseau représentent les variables des quatre niveaux décrivant ci-dessus et les arcs décrivent les dépendances entre ces nœuds. La figure (Fig.4.6) illustre la première étape consistant à modéliser qualitativement le problème et à déterminer les influences existantes entre les variables.

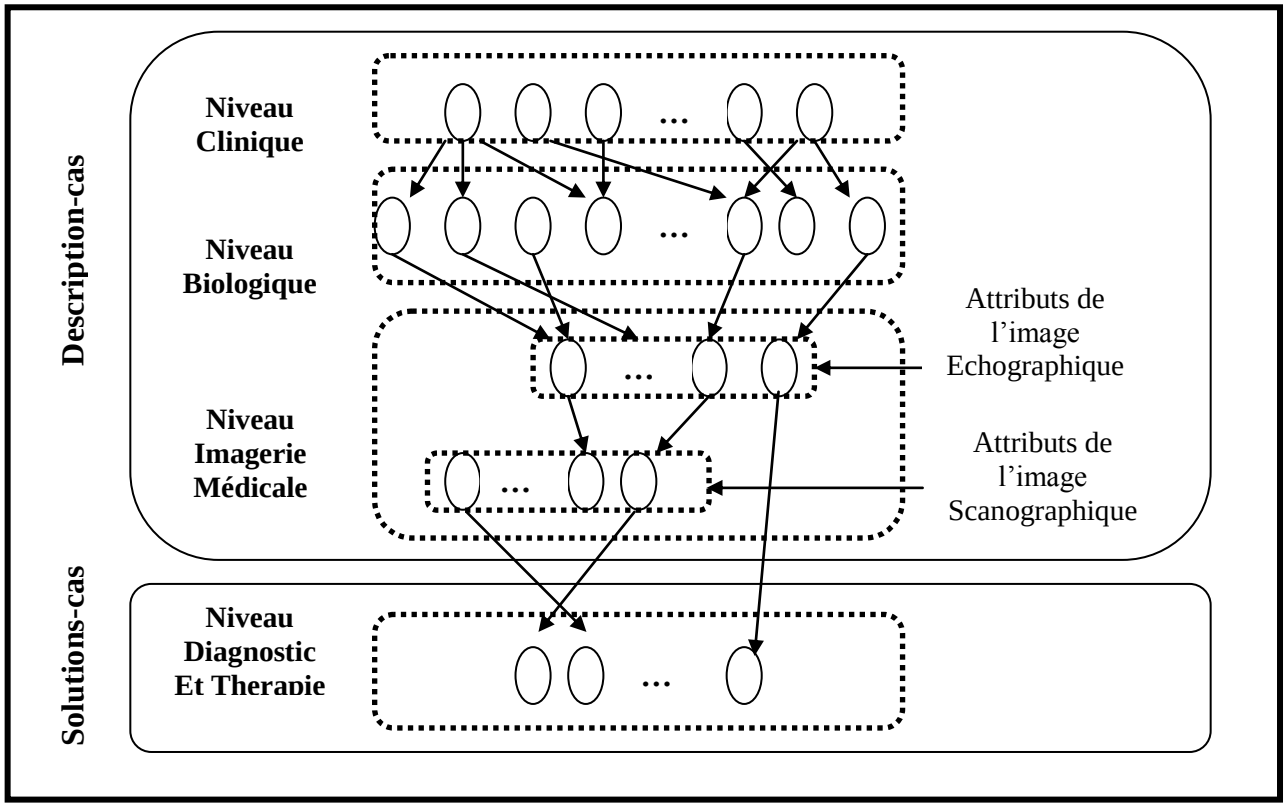



Fig.4.6. Modélisation de la base de cas par un réseau bayésien

### 5.2. Description d'un cas du système CBR

Le cas est décrit en termes de dimension attributs. Ces attributs sont des données cliniques, des données biologiques, des caractéristiques extraites à partir de l'image du foie, et de diagnostic conclu avec sa thérapie. Par définition un cas est une entité qui contient des diverses informations représentant un domaine particulier. Un cas contient les données empiriques décrivant l'expérience acquise dans la résolution d'une situation précise.

Dans notre étude, les attributs du cas sont des données cliniques, des données biologiques, des caractéristiques extraites à partir de l'image échographiques et scannographiques du foie [DJE 04], et du diagnostic établi.

Un cas comporte deux parties, à savoir (Fig.4.7):

- La description du problème : les descripteurs cliniques, biologiques, et les images

échographiques et scanographiques du foie.

- La description de la solution : le diagnostic et la thérapie.

C'est l'information plus ou moins standardisée contenue dans les comptes-rendus qui a servi de base à la conception des cas.

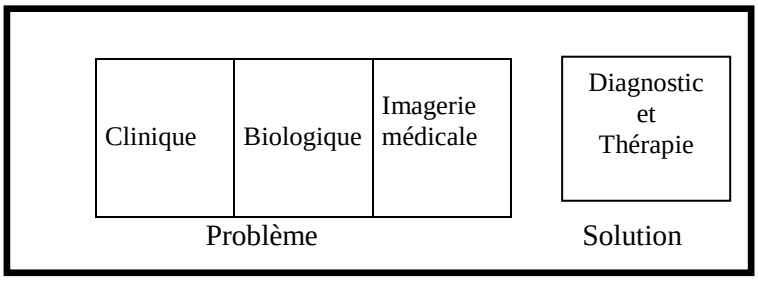



Fig.4.7. Description d'un cas

Dans cette application, le cas est représenté par un ensemble d'attributs qui sont décrit dans la figure (Fig.4.8) :

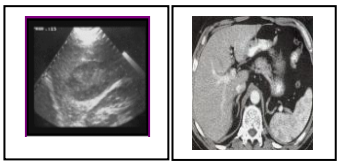
**Description d'un cas:**

**Signes clinique :**
Douleur=1
Ictère= 0
Tabac=1

**Signes biologiques**
……..

**Imagerie médicale**

Attributs de l'image échographique
Taille=…
Homogénéité=…



Attributs de l'image scannographique
Taille=620
Densité=125
Angle aigu=…
Homogénéité=… ……

**Solution (diagnostic-thérapie) :**
Cirrhose-Biopsie

Fig.4.8. Exemple d'un cas du système RàPC

### 5.3. Architecture de la base de cas

Le système proposé est constitué d'un sous système de traitement d'image (traitement des images échographiques et scanographiques de foie). Ce dernier passe par plusieurs phases [DJE 04]:

- Une phase de prétraitement suit l'acquisition et la numérisation de l'image du foie. Elle consiste à éliminer une quantité importante de bruits.
- Une phase de segmentation qui consiste à isoler, les uns des autres, les éléments

présents dans l'image.

- Une phase d'analyse qui sert à extraire les indices (taille, périmètre, densité,…..) représentants l'image du foie.

## 6. La phase de la remémoration

La base de cas de ce système est une collection de vecteurs de cas, notés $C=\{(C_1,P_1), (C_2,P_2),\dots,(C_l,P_l)\}$. Chaque cas est décrit par deux vecteurs $C_k$ et $P_k$, tel que $C_k=(C_{k1},C_{k2},\dots C_{kn})$, $C_{ki}\in R$ une combinaison des valeurs numériques associées à chaque attribut du cas k et $P_k=(P_{k1},P_{k2},\dots P_{kn})$, $P_{ki}$ [0,1] est un vecteur de probabilité qui contient des probabilités approximatives aux attributs du cas. Et, comme la construction des probabilités dans un réseau bayésien est une étape souvent difficile, où elle nécessite la plus part du temps des avis des experts et afin d'optimiser l'étape de calcul, nous considérons la mesure de similarité comme un modèle log linéaire [DJE 12a].

Cette phase de remémoration passe par trois processus (Fig.4.9) qui seront décrits dans les sections suivantes :

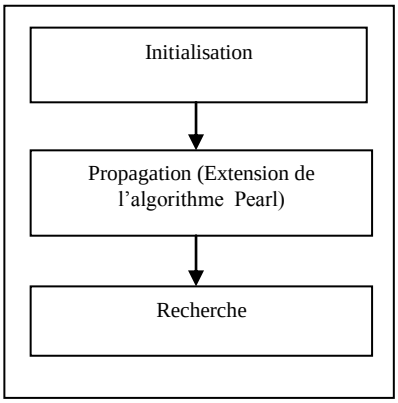



Fig.4.9. Etapes de la phase de remémoration

### 6.1. Processus d'initialisation

Ce processus permet d'initialiser les variables du réseau et d'attribuer à chaque niveau i un poids $\lambda_i$. Ces variables sont les probabilités marginales à priori p(x) du premier niveau (le niveau clinique) et les probabilités conditionnelles des autres niveaux. p(S|D) est la probabilité postériori de symptômes k{1..k} des niveaux cliniques, biologiques et l'imagerie médicales qui sont donnés par l'observation x.

### 6.2. Processus de propagation (Extension de l'algorithme Pearl)

L'extension de l'algorithme Pearl réside dans l'intégration du poids $\lambda_i$ à chaque niveau i. Durant le processus de propagation, chaque probabilité du nœud sera multiplie par son poids correspondant. Une fois les variables sont initialisées par un ensemble de probabilités, le processus de propagation de message s'effectue à chaque niveau de réseau jusqu'au niveau inferieur (fils). Ceci est fait par un passage de message π qui permet d'entrainer pour chaque variable une mise à jour de sa probabilité. Ce message π venant du père vers les fils est noté pour chaque variable comme une probabilité actuelle [PEA 88].

Le message π des variables du premier niveau (clinique) est calculé selon la formule suivante :

$$\pi(x) = \lambda_i \, p(x) \qquad (2)$$

où :

π (x) est le message venant de x.

$\lambda_i$ est le poids du niveau i.

p(x) est la probabilité marginale.

Les variables de deuxième niveau, reçoivent ce message et mettent alors à jour ses probabilités en entrainant le message π dans le calcul des probabilités actuelles. Un nouveau message π est obtenu selon la formule suivante :

$$\pi(X=x) = \lambda_i \sum z \, P(X=x|Z=z) \, \pi x \, (Z=z) \qquad (3)$$

où:

π (x) est le message venant de x.

$\lambda_i$ est le poids du niveau i.

Z est le parent.

P (X=x|Z=z) est la probabilité de X sachant ses parents.

π x (Z=z) est le message venant de Z à son fils x.

Le même principe est effectué sur le troisième niveau. Pour les variables du dernier niveau, il n'y a pas de message π à envoyer. Lorsque toutes les variables du réseau mettent à jour ses probabilités à la lumière du message π reçu, le processus de propagation se termine. Le résultat obtenu au dernier niveau permet de lancer le processus de recherche qui va fournir le résultat final du diagnostic.

### 6.3. Processus de recherche

La recherche s'effectue selon l'évaluation de la ressemblance entre le nouveau cas et les cas déjà connus dans la base de cas. Soit C le nouveau cas décrit par un ensemble d'observations X={$X_1$, $X_2$,..,$X_n$} qui sont des signes cliniques, biologiques et des attributs de l'imagerie médicale du foie (l'échographie et le scanner TDM[5]). Cette phase consiste à déterminer le diagnostic associé à ce nouveau cas en se basant sur les probabilités (obtenus par le modèle log-linéaire) comme des mesures de similarité. La recherche du cas similaire évalué par la quantité p(D|S) est la probabilité que les descripteurs de cas S= $S_1$, $S_2$,.., Sj, sachant le diagnostic de cas D= $D_1$, $D_2$,.., $D_j$. On introduit l'ensemble de l'hiérarchie, puis en pratique, le résultat du processus de propagation est le diagnostic associé au nouveau cas [DJE 11a].

### 6.3.1. L'utilisation du model Log linéaire

Dans cette étude, nous avons pondéré chaque niveau du réseau par un poids $\lambda_i$ : le niveau clinique a un poids $\lambda_1$, le niveau biologique a un poids $\lambda_2$, le niveau imagerie médicale échographie a un poids $\lambda_3$ et le niveau imagerie médicale TDM a un poids $\lambda_4$. Les poids accordés à chaque niveau, correspondant à leur pondération dans la combinaison log-linéaire, constituent les paramètres du modèle. Le principe de la phase de remémoration est de représenter la mesure de similarité par une combinaison log-linéaire de plusieurs probabilités P(D|S) pondérées par des poids $\lambda_i$ (Extended Pearl). Le facteur de pondération $\lambda_i$ est utile pour introduire le maître symptôme de chaque niveau [DJE 11b]. La phase de recherche cherche à maximiser l'expression suivante (Fig.4.10):

$$argmax \sum_{i=1}^{n} (\lambda_i \log P(D = d|S) \qquad (4)$$

où

$\lambda_i$ : le poids du niveau i.

D : le diagnostic associé au cas le plus similaire.

S : les symptômes.

P(D=d|S) est la probabilité conditionnelle.

En effet, on recherche la maladie ayant la probabilité maximum sachant les probabilités et les poids $\lambda_i$.

[5] TDM : *Scanner ou Tomodensitométrie*

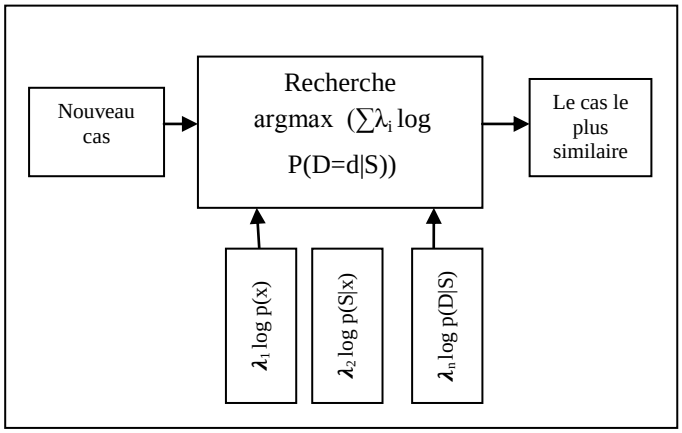



Fig.4.10. Modélisation de la mesure de similarité par un modèle Log-linéaire

### 6.3.2. L'algorithme de la remémoration proposé

Ce paragraphe résume la phase de remémoration par l'algorithme 2. Cet algorithme commence par la sélection des signes cliniques, biologique, et les attributs d'imagerie médicale. Une fois, ces informations sont introduites, chaque variable du réseau bayésien envoie un message de type π, contenant sa probabilité actuelle, à tous ses fils sauf les variables du dernier niveau qui n'ont pas de fils. A la réception d'un tel message, les variables doivent remettre à jour ses nouvelles probabilités à la lumière de cette nouvelle information (le message π) puis elles envoient à leur tour des messages π à ses fils. A chaque envoi, nous calculons la somme du logarithme de chaque probabilité multiplié par son poids correspondant. A la fin de ce processus, toutes les feuilles (variables du dernier niveau) remettent à jour leurs probabilités afin d'obtenir leurs probabilités à posteriori. Enfin, l'algorithme cherche la somme maximale du dernier niveau qui sera le résultat du diagnostic du cas remémoré [DJE 11c].

**Algorithme 2.** Algorithme de la remémoration proposé.

```
Entrée : BN
λi: le poids du niveau i
i : le numéro du niveau
j : le numéro du symptôme
N : le nombre des niveaux dans le BN
M : le nombre de nœuds actifs du niveau i
Nouveau cas (ensemble de symptômes S)
Sortie : solution source : max (Somi)
Début
Pour i allant de 1 à N Faire
     Pour j allant de 1 à M Faire
               Début
               Propagation des nœuds symptômes S ;
               Somi := Somi + λi log P(D=d|Sj) ;
               Fin;
Fin;
D:=max (Somi); // D est le diagnostic associé au cas le
plus similaire
Fin.
```

## 7. La phase d'adaptation

La phase d'adaptation sert à modifier les solutions des cas similaires pour adapter un problème cible. S'il n'y a aucune différence importante entre le cas cible et le cas similaire, un transfert de solution est suffisant. Parfois peu de substitutions sont exigées, mais habituellement, l'adaptation est un processus compliqué. Puisque l'adaptation est encore plus difficile en médecine, nous voulons élaborer des problèmes d'adaptation médicaux et nous espérons montrer des possibilités sur la façon de les résoudre.

Le processus d'adaptation dans la figure (Fig.4.11) commence par la phase de la remémoration (basé sur les différents symptômes de la description du cas cible et la description du cas remémoré). Cette question est utilisée pour retrouver l'adaptation du cas le plus similaire. La similarité entre le cas cible et le cas remémoré est de nouveau évaluée. Si la mesure d'adaptation est supérieure au seuil 3 (voir le paragraphe suivant), la solution offerte par le cas remémoré est directement réutilisée. Cela signifie que la solution du cas cible de diagnostic correspond à la solution du cas remémoré. Si la mesure d'adaptation est inférieure ou égale au seuil 3 alors on va appliquer des règles thérapeutiques ou la biopsie du foie. Le seuil d'adaptation sera défini dans le paragraphe suivant.

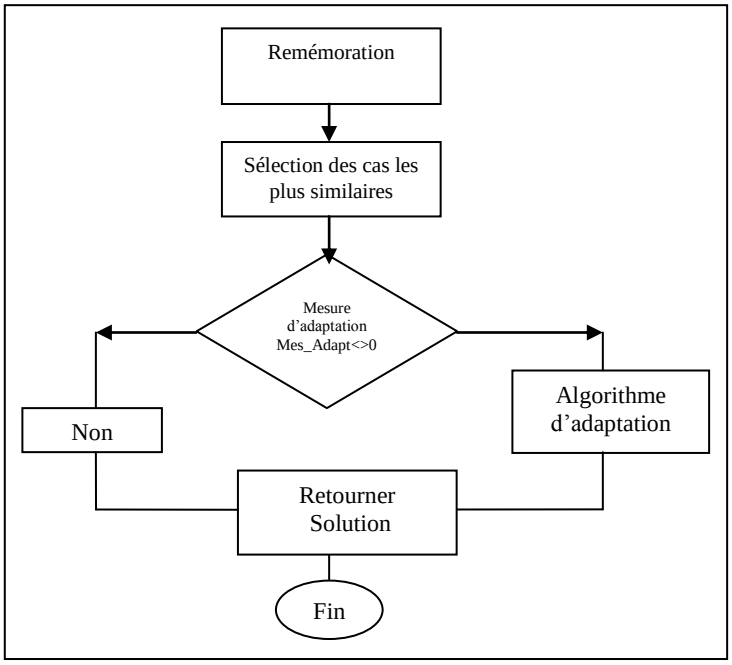



Fig.4.11. Processus d'adaptation

*Définition de la valeur du seuil d'adaptation*

Dans le paragraphe précédent, nous avons utilisé un seuil pour la mesure d'adaptation, sa valeur a été déterminée selon le degré d'importance du niveau d'imagerie médical pour le diagnostic des pathologies hépatiques. S'il existe au moins un seul maître symptôme[6] du niveau imagerie médical, cela implique que : *Adapt_Meas* = 0 x $\lambda_1$+0 x $\lambda_2$+1 x $\lambda_3$=1 x 3=3. La solution d'adaptation sera la solution du cas remémoré ou la solution de sa classe correspondante[7] [DJE 12b][DJE 12c].

### 7.1. Définition de la mesure d'adaptation

Lorsque le nombre de maîtres symptômes pour chaque niveau du cas remémoré est déterminé, ce nombre sera multiplié par un facteur de pondération. Alors la somme du maître symptôme de tous les attributs est calculée pour fournir une mesure d'adaptation du cas remémoré à celui du cas cible. Cette mesure est définie par la formule suivante [DJE 12b][DJE 12c]:

[6] Maître symptôme : est le facteur cause d'une maladie.

[7] Classe correspondante : les maladies du foie sont réparties en classe, voir chapitre 4 (section 4.2).

$$Adapt\_Meas(T, R) = \sum_{i=1}^{n} (\lambda_i * NB\_MA) \quad (5)$$

Où :

T : cas cible

R : cas remémoré

i : le nombre du niveau allant de 1 à n

MA : Maître Symptôme est le symptôme qui a une probabilité maximale.

*NB_MA* : le nombre de maîtres symptômes pour chaque niveau i

$\lambda_i$: le poids de l'attribut i

### 7.2. L'algorithme d'adaptation proposé

Soit e un examen effectué sur un patient donné, avec $e_{ij}$ l'ensemble des symptômes cliniques, biologiques et imagerie médicale, tel que i est le numéro du niveau et j est le numéro du symptôme. Après l'application de l'algorithme de la remémoration, on obtient le cas remémoré. En suite, on applique la mesure d'adaptation *Adapt_Meas* [DJE 12b] [DJE 12e]:

Si Mes_Adapt>3 alors la solution du cas cible sera la solution du cas remémoré

Sinon la solution du cas cible sera la solution de la classe correspondante au cas remémoré.

Si Mes_Adapt=0 alors « Pas d'adaptation ».

L'algorithme d'adaptation (Algorithme 3) décrit ci-dessous résume les étapes d'adaptation.

**Algorithme 3.** Algorithme d'adaptation proposé.

Entrée :
BN : base de cas (Ps : partie problème +Ss : partie solution)
i : le numéro du niveau
j : le numéro du symptôme
N : le nombre de niveaux du BN
M : le nombre de nœuds actifs par niveau i
dp : sont les descripteurs de la partie problème du cas remémoré
Sortie : trouver la solution du cas cible
**Début**
Créer une liste contenant les maîtres symptômes de chaque maladie
**Pour** i allant de 1 à N **Faire**
**Pour** j allant de 1 à M **Faire**
Début
**Si** ($d_1$ n'est pas un maître symptôme clinique) et ($d_2$ n'est pas un maître symptôme biologique) et ($d_3$ n'est pas un maître symptôme imagerie médicale) **alors *pas d'adaptation*** ;
**Si** ($d_1$ n'est pas un maître symptôme clinique) et ($d_2$ n'est pas un maître symptôme biologique) et ($d_3$ est un maître symptôme imagerie médicale) **alors *copier la solution de la classe correspondante au cas remémoré***;
**Si** ($d_1$ n'est pas un maître symptôme clinique) et ($d_2$ est un maître symptôme biologique) et ($d_3$ n'est pas un maître symptôme imagerie médicale) **alors *copier la solution de la classe correspondante au cas remémoré***;
**Si** ($d_1$ n'est pas un maître symptôme clinique) et ($d_2$ est un maître symptôme biologique) et ($d_3$ est un maître symptôme imagerie médicale) **alors *copier la solution remémoré*** ;
**Si (**$d_1$ est un maître symptôme clinique) et ($d_2$ n'est pas un maître symptôme biologique) et ($d_3$ n'est pas un maître symptôme imagerie médicale) **alors «Traitement médical »**;
**Si** ($d_1$ est un maître symptôme clinique) et ($d_2$ n'est pas un maître symptôme biologique) et ($d_3$ est un maître symptôme imagerie médicale) **alors *copier la solution remémoré***;
**Si** ($d_1$ est un maître symptôme clinique) et ($d_2$ est un maître symptôme biologique) et ($d_3$ n'est pas un maître symptôme imagerie médicale) **alors *copier la solution de la classe correspondante au cas remémoré***;
**Si** ($d_1$ est un maître symptôme clinique) et ($d_2$ est un maître symptôme biologique) et ($d_3$ est un maître symptôme imagerie médicale) **alors *copier la solution remémoré***;
Fin ;
Fin.

Cet algorithme traite la phase d'adaptation qui suit la phase de la remémoration. Il se base sur l'attribut « maître symptôme » comme une mesure d'adaptation. A travers ce dernier, nous pouvons adapter les solutions des cas proches dans la base de cas. Dans notre approche, une fois qu'un cas remémoré est trouvé, il faut chercher tous les maîtres symptômes correspondants à ce cas et faire une comparaison niveau par niveau (clinique, biologique et imagerie médical) entre le cas remémoré et le cas cible. S'ils sont très proches, copie la solution du cas remémoré au cas cible. Si le système n'arrive pas à trouver un problème proche dans la base de cas, il va déclencher soit des règles thérapeutiques soit la biopsie du foie. Le résultat obtenu sera une solution au problème cible.

**Exemple illustratif :**

**Exemple 01 : « copier la solution du cas remémoré »**

Soit le cas cible A décrivant les trois niveaux du réseau bayésien et le cas source remémoré (Tab.4.4).

| | Descripteurs | Cas cible A | Cas remémoré |
|---|---|---|---|
| Description du problème | Ictère | | 0.1 |
| | Hépatomégalie | | |
| | Hypertension portale | 0.1 | 0.3 |
| | Alcool | 0.8 | **0.9** |
| | Insuffisance rénale | 0.8 | **0.9** |
| | Hépatite viral | 0.1 | |
| | AgHg | | |
| | Albuminemie, Phosphatase | - | 0.1 |
| | Bilirubine | | 0.1 |
| | Hyperglobuline | 0.69 | **0.7** |
| | Taille du foie | | |
| | Homogénéité | 0.7 | **0.9** |
| | Atrophie | | **0.8** |
| | Taille de lésion | | |
| Description du solution | | ? | Cirrhose |

Tab.4.4. La solution d'adaptation du cas cible A

*Solution d'adaptation :*

On voit que pour chaque niveau, il y a au moins un maître symptôme, ce qui implique de copier la solution du cas source.

-pour le niveau clinique : les signes « Alcool » et « Insuffisasnce rénale » sont des maîtres symptômes puisque ils portent des fortes probabilités.

- pour le niveau biologique : le descripteur « Hyperglobuline » est un maître symptôme.

-pour le niveau Imagerie médicale : l'image échographique du foie est de taille normale mais elle est hétérogène et l'image scannographique est non homogène et hyperdense.

Nous appliquons la mesure d'adaptation proposée :

$Adapt\_Meas(T, R) = \sum_{i=1}^{n}(\lambda_i * NB\_MA)$

*Adapt_Meas*(*T*, R) =2*1+1*2+1*3 =7 >3

La solution de l'adaptation est de copier la solution du cas remémoré : la maladie Cirrhose.

**Exemple 02** : **« Copier la solution de la classe correspondante au cas remémoré»**

Soit le cas cible B décrivant les trois niveaux du réseau bayésien et le cas source remémoré (Tab.4.5).

| | Descripteurs | Cas cible B | Cas remémoré |
|---|---|---|---|
| Description du problème | Découverte Fortuite | 0.1 | 0.1 |
| | Ictère | | 0.5 |
| | Hépatite b/c | 0.3 | **0.7** |
| | Douleur | 0.5 | 0.5 |
| | Alcool | | 0.3 |
| | AgHg | | **0.7** |
| | Phosphatase | | 0.5 |
| | Bilirubine | | 0.5 |
| | Alpha foeto protéine | 0.4 | **0.7** |
| | Taille du foie | | 0.5 |
| | Homogénéité | 0.5 | 0.6 |
| | Taille de lésion | 0.7 | **0.7** |
| | Densité | | |
| Description du solution | | ? | Carcinome hépatocellulaire |

Tab.4.5. La solution d'adaptation du cas cible B

*Solution d'adaptation :*

On voit que le niveau clinique et le niveau biologique, il n'y a pas un maître symptôme mais il existe qu'un seul maître symptôme de l'imagerie médicale, ce qui implique que la solution d'adaptation consiste à copier la solution de la classe correspondante au cas remémoré avec l'application des règles thérapeutiques ou la biopsie du foie.

Nous appliquons la mesure d'adaptation proposée :

*Adapt_Meas*(*T*, R) =$\sum_{i=1}^{n}(\lambda_i * NB\_MA)$

*Adapt_Meas* (*T*, R) =0*1+0*2+1*3 =3 ≤3

La solution de l'adaptation est la solution de la classe correspondante à la maladie Carcinome hépatocellulaire : la classe Tumeur Maligne.

Exemple d'application des règles thérapeutiques : ces règles dépendent de la taille de la lésion.

*Chirurgie :*

Elle est effectuée lorsque la lésion est unique ou lorsque les lésions sont multiples pouvant être volumineuses;

*Transplantation :*

Elle est indiquée pour des lésions de petite taille et en petit nombre (moins de trois nodules) dont la résection n'est pas possible du fait de la localisation ou de l'insuffisance hépatocellulaire;

*Traitement percutanés :*

Ils sont réalisés lorsque la lésion est unique, de 3 à 5cm au maximum, ou lorsqu'il existe deux nodules, repérables et ponction nables en échographie qu'est la technique habituellement utilisée pour réaliser le geste. Les contre-indications sont les troubles de l'hémostase et la présence d'ascite;

*Chimioembolisation :*

L'injection intra-artérielle sélective dans l'artère hépatique d'un mélange de chimiothérapie (adriamycine) et de Lipiodol ultrafluide (LUF) suivie d'une embolisation en amont peut être réalisée, mais son bénéfice n'a pas été montré sur la survie. Une toxicité directe sur l'artère est fréquente. Elle est réservée le plus souvent aux lésions inopérables : CHC multifical ou diffus.

**Exemple 03 : « Pas d'adaptation »**

Si dans les trois niveaux, il n'a pas un descripteur qui représente un maître symptôme. C'est-à-dire il n'existe pas un cas plus proche au cas cible, on doit intervenir l'avis de l'expert « médecin » pour ne pas proposer des solutions non pertinentes au problème cible.

*Adapt_Meas*(*T*, R) =$\sum_{i=1}^{n}(\lambda_i * NB_{MA})$

*Adapt_Meas*(*T*, R) =0*1+0*2+0*3 =0

## 8. Discussion

Nous avons modélisé la mesure de similarité par un modèle log linéaire qui consiste à exprimer des logarithmes des probabilités. Cette modélisation permet non seulement, d'optimiser les calculs de probabilités nécessaires à l'inférence mais aussi d'avoir l'avantage d'être hiérarchique. Cette mesure de similarité a pour rôle de déterminer le cas le plus similaire en se basant sur le poids du niveau du réseau et le logarithme des probabilités. Elle est une combinaison log linéaire de plusieurs probabilités pondérées par des poids $\lambda_i$.

Notons que le poids (le facteur de pondération) est utile pour introduire le maître symptôme parmi l'ensemble des symptômes. Le logarithme est une fonction de simplification de calcul et en particulier elle permet de remplacer des multiplications par des sommes afin de simplifier la lourde tâche des calculateurs (donc optimiser les calculs). Ce qui rend la remémoration plus efficace (rapide).

Pour la phase d'adaptation, nous avons montré à travers les trois exemples que les cas les plus similaires au cas cible ne sont pas forcement ceux qu'on choisit lors de la phase d'adaptation. Et la mesure d'adaptation *Adapt_Meas* a l'avantage de faciliter l'adaptation du cas remémoré en se basant sur deux paramètres essentielles qui sont : NB_*MA* : le nombre de maître symptôme pour chaque niveau i et $\lambda_i$: le poid de l'attribut i. Plus le nombre de maîtres symptôme augmente avec un poids élevé plus l'adaptation sera facile. C'est-à-dire que les maîtres symptômes du niveau imagerie médical jouent un rôle très important pour rendre la phase d'adaptation plus facile et plus flexible.

## 9. Conclusion

Dans ce chapitre, nous avons présenté une modélisation statistique de la phase de remémoration et d'adaptation d'un RàPC dans le but d'aide au diagnostic médical, en mettant l'accent sur le formalisme des réseaux bayésiens. L'approche décrite montre le rôle de cette modélisation dans le domaine médicale. Comme les connaissances de notre domaine sont incertaines, ceci nous a poussés à étudier un modèle probabiliste de mémoire de cas pour l'aide au diagnostic des pathologies hépatiques. Afin de ne pas limiter notre modèle sur le plan de modélisation de connaissances, nous avons élaborés deux phases : la phase de remémoration et la phase d'adaptation. Dans la phase de remémoration, une extension de l'algorithme Pearl a été proposée avec une mesure de similarité Log-linéaire afin de permettre une recherche efficace orientée vers l'adaptabilité des solutions des cas sources. La phase d'adaptation est une étape de certains systèmes de raisonnement à partir de cas qui consiste à

modifier un cas source pour qu'il réponde à une nouvelle situation, le cas cible. Dans ce cadre, nous avons proposé un algorithme d'adaptation qui s'appui sur les paramètres du réseau bayésien représentant la base de cas. Un exemple d'aide au diagnostic des pathologies hépatiques est illustré pour l'algorithme proposé. En effet, ces deux phases sont unifiées entre elles. Cette dépendance est définie par deux mesures : mesure de similarité et mesure d'adaptation. Elle a pour objectif de sélectionner le cas le plus facilement adaptable pour la phase d'adaptation.

La validation de notre approche est présentée au chapitre suivant.

# Chapitre 5

# Expérimentation

## Sommaire

# 1. Introduction

Comme tout modèle qui doit être expérimenté, le présent chapitre constitue un cadre d'expérimentation et d'argumentation du chapitre précédent.

Nous avons implémenté les deux phases la remémoration et l'adaptation, en établissant une base de cas de 19 maladies les plus fréquentes du foie. Cette application a été conçue dans l'environnement Visuel C++, Ultimate 2010. Nous nous intéressons aussi à la validation de la phase de la remémoration par les algorithmes : Pearl, Extended Pearl et JLO à travers lesquels nous avons établis une étude comparative.

# 2. Validation des phases de remémoration et d'adaptation

Après avoir défini formellement notre approche, il est nécessaire de la tester afin de la valider.

## 2.1. Base de tests

Pour valider les deux phases, nous avons réalisé des tests sur une base de 19 maladies les plus fréquentes du foie. La base de test est constituée de 116 cas obtenus soit par un changement de symptômes soit par un changement de probabilités du réseau bayésien. Elle est répartit en 6 classes suivantes [DJE 12c]:

- La 1ère classe : obtenue par un changement aléatoire d'arc du réseau bayésien.
- La 2ère classe : obtenue par un changement d'une probabilité d'un symptôme avec un changement de son type correspondant (maître ou non maître symptôme).
- La 3ère classe : obtenue par un changement de deux probabilités avec ses types correspondants.
- La 4ère classe : obtenue par un changement d'un seul symptôme.
- La 5ère classe : obtenue par un changement de deux symptômes.
- La 6ère classe : représente la même structure de la base de cas (réseau bayésien).

La figure (Fig.5.1) montre un aperçu de la base de tests.

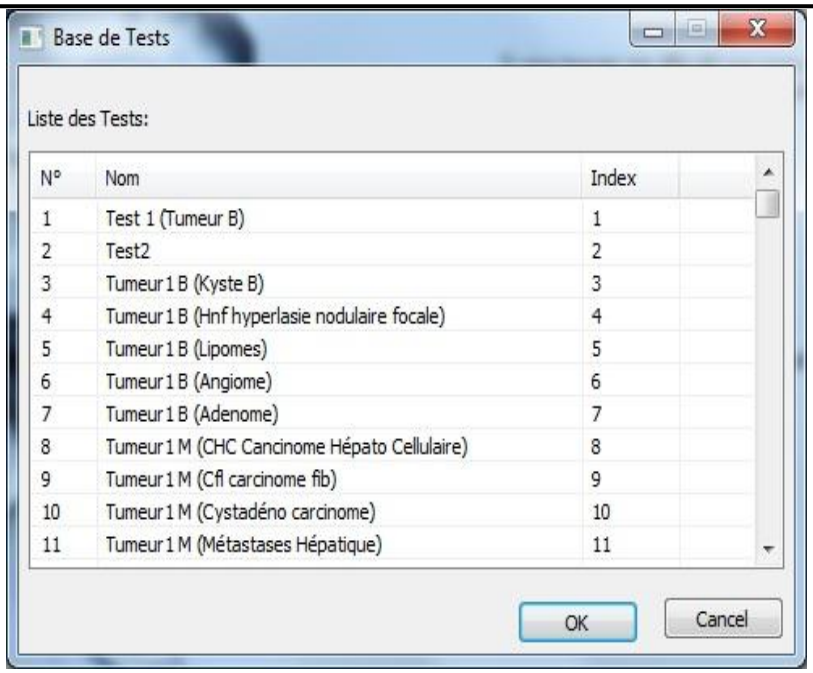



Fig.5.1. La base de test

## 2.2. La phase de remémoration

Pour la phase de remémoration, nous avons comparé les deux algorithmes : l'algorithme Pearl et notre algorithme de remémoration proposé. La figure (Fig.5.2) montre que l'évolution du taux de similarité de l'algorithme Extended Pearl est meilleure par rapport à l'algorithme Pearl. Ce qui implique que l'utilisation de modèle log linéaire a influencé sur le calcul des probabilités et à optimiser la phase de remémoration.

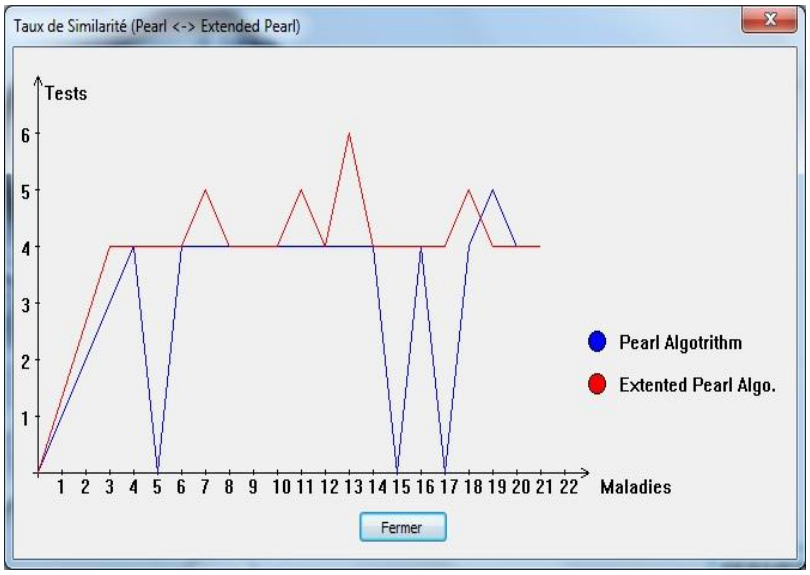



Fig.5.2. Evolution de la mesure de similarité de l'algorithme Pearl et l'algorithme Extended Pearl

## 2.3. La phase d'adaptation

Le but de protocole d'évaluation de la phase d'adaptation est de prouver la faisabilité de celle-ci en calculant la précision de la base de cas. Afin de réaliser cette évaluation, nous appliquons la phase de la remémoration proposée décrite dans la section précédente, à travers le cas remémoré obtenu, nous appliquons la mesure d'adaptation pour établir une solution au cas cible que nous avons élaboré dans le chapitre 4. Notre méthode a été comparée avec une méthode qui n'utilise pas d'adaptation. Les résultats issus sont illustrés par la figure (Fig.5.3).

Par ces résultats, nous déduisons que la méthode d'adaptation proposée a réussi sur un sous ensemble de la base de cas, on voit sur la courbe rouge qu'on perd de la précision si on est proche au cas de changement de type de symptôme de la deuxième classe et nous remarquons aussi que l'adaptation des cas cibles a une évolution non constante. Cependant, nous constatons que les résultats obtenus avec et sans adaptation ne sont pas vraiment mitigés. Notre méthode d'adaptation donne des résultats pas très satisfaisants mais ils sont encourageants.

Cela nous a permis de constater que les cas les plus similaires ne sont pas forcément les cas choisis pour la phase d'adaptation. Pour cela nous envisageons en perspective d'implémenter cette méthode sur un cas réel de grande taille. Ces résultats montrent que notre méthode d'adaptation a assuré une adaptation acceptable, sur l'ensemble des cas de la base de test, par rapport à la méthode sans adaptation.

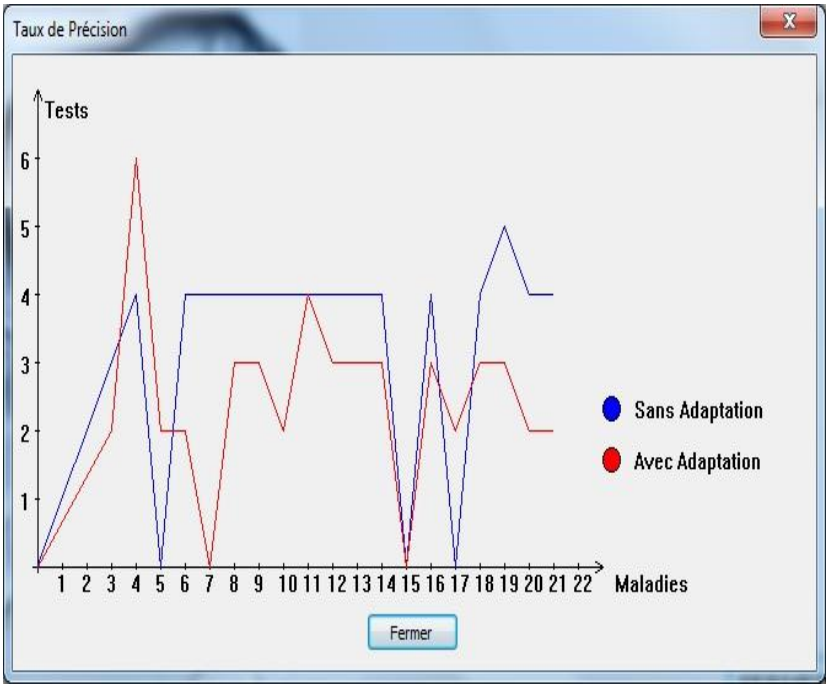



Fig.5.3. Evolution de la précision sans et avec l'adaptation

D'après les résultats obtenus, nous conclurons que le RàPC peut raisonner à partir d'un nombre restreint de cas. En médecine l'adaptation se diffère d'un cas à un autre puisque elle est encore plus difficile car nous voulons élaborer des problèmes d'adaptation médicaux et nous espérons montrer des possibilités sur la façon de les résoudre.

## 3. Remémoration par l'algorithme Pearl

Cette phase de remémoration de cas similaires dépend essentiellement de la représentation des cas, de la structure de base de cas, des mesures de similarité. Pour permettre de comparer les cas les uns avec les autres, il faut pouvoir comparer leurs valeurs d'attributs de façon à établir à quels points ces valeurs sont proches. Chaque attribut doit donc être typé. C'est la connaissance du type qui permet de connaître les opérations de comparaison licites et par là d'établir des similarités. Notre système adopte les réseaux bayésiens et il repose sur l'algorithme d'inférence exacte Pearl qui a été élaboré au chapitre 4.

Le système demande en entrée de sélectionner les signes cliniques, biologiques et les attributs de l'imagerie médicale (Fig.5.4). Après entrer de ces informations, chaque variable du réseau bayésien envoie un message de type π à tous ses fils sauf les variables du dernier niveau qui n'ont pas de fils. A la réception d'un tel message, les variables doivent remettre à jour ses nouvelles probabilités à la lumière de cette nouvelle information (le message π) puis elles envoient à leur tour des messages π à ses fils. A la fin de ce processus, toutes les feuilles (variables du dernier niveau) remettent à jour leurs probabilités afin d'obtenir leurs probabilités à posteriori. Le système cherche la variable du dernier niveau ayant la probabilité maximale (comme mesure de similarité) sachant qu'elle représente le résultat du diagnostic [DJE 07].

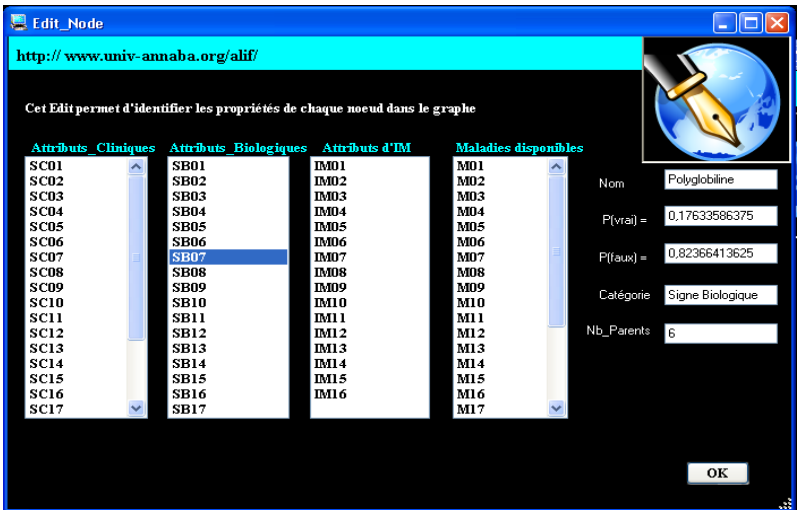



Fig.5.4. Attributs du cas cible

Nous avons implémenté un réseau bayésien de 81 nœuds, 56 nœuds représentant les signes de trois niveaux (clinique, biologique et imagerie médicale) et 25 nœuds représentant le niveau diagnostic et sa thérapie (Fig.5.5). Nous avons obtenu un taux de similarité de 79.68% (Fig.5.6) [DJE 12d].

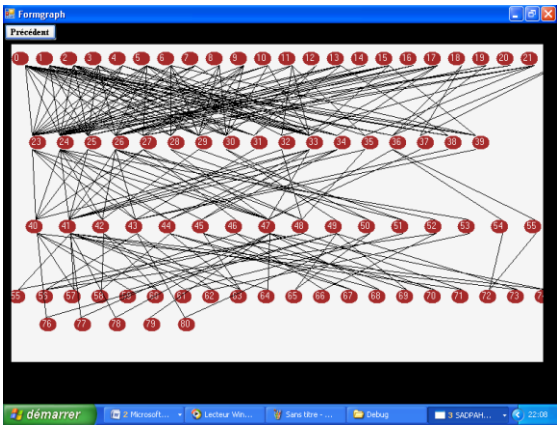



Fig.5.5. Le réseau associé à la base de cas

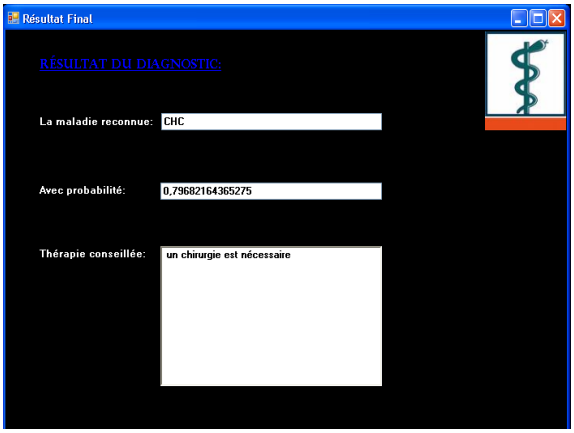



Fig.5.6. Résultat de diagnostic

## 4. Remémoration par l'algorithme de l'arbre de Jonction JLO

L'algorithme de l'arbre de jonction est dit JLO, du nom de ses auteurs: [JEN 90] (définit dans le chapitre 2, section 8.1.2). Il s'applique à des réseaux ne comprenant que des variables à valeurs discrètes. L'algorithme comporte deux phases: une phase de construction et une phase de propagation. La phase de construction permet de transformer le réseau bayésien initial en un arbre de jonction, dont les nœuds sont des cliques de nœuds du réseau initial.

Cette transformation se fait en trois étapes : la moralisation du graphe, la triangulation du graphe et la création d'un arbre appelé arbre de Jonction qui est constitué de 57 cliques (Fig.5.7). Une fois les étapes de construction de l'arbre de jonction sont effectuées, nous passons à la phase de diagnostic du nouveau cas (Fig.5.8)

La solution reconnue est celle qui a la vraisemblance $P(C_i|e)$ maximale.

où :

e : l'ensemble des observations correspondant au nouveau cas.

$C_i$ : la clique correspond à la solution.

Pour montrer la capacité de la modélisation de la base de cas par un réseau bayésien, nous avons réalisé un test sur une base de cinq pathologies. Ce test est implémenté sous Builder C++ version 6.0 [DJE 06].

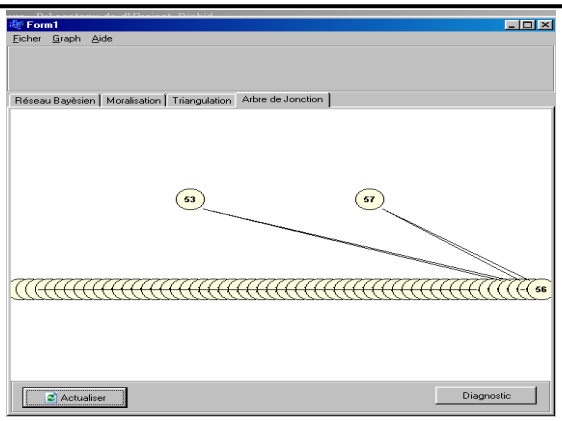



Fig.5.7. L'arbre de Jonction

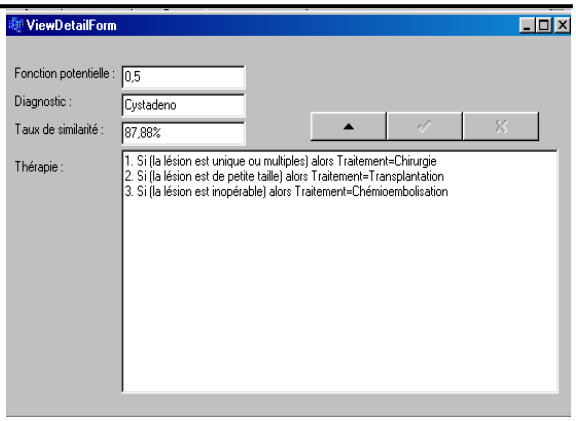



Fig.5.8. Classification

Le réseau est composé de 21 nœuds:

- 16 nœuds représentant les signes de trois niveaux (clinique, biologique, imagerie médicale)
- 5 nœuds représentant le niveau diagnostic et thérapie.

Nous retenons ici l'ensemble des 21 variables décrites dans le tableau Tab.5.1.

| | Variable | Description de la variable |
|---|---|---|
| Description du problème | 1 | Découverte Fortuite |
| | 2 | Ictère |
| | 3 | Hépatomégalie |
| | 4 | Douleur |
| | 5 | Alcool |
| | 6 | Hépatite virale |
| | 7 | AgHg |
| | 8 | Phosphatase |
| | 9 | Bilirubine |
| | 10 | Sérologie |
| | 11 | Hyperglobuline |
| | 12 | Taille du foie |
| | 13 | Homogénéité |
| | 14 | Atrophie |
| | 15 | Taille de lésion |
| | 16 | Densité |
| Description du solution | 17 | Hyperplasie nodulaire Focale |
| | 18 | Amylose |
| | 19 | Cystadéno |
| | 20 | Kyste Hépatique |
| | 21 | Cirrhose |

Tab.5.1. Les descripteurs de cas

Nous remarquons que l'algorithme JLO a donné des résultats encourageants, puisque, sur une base de 5 cas nous avons obtenus un taux de similarité de 87.88%. Les résultats montrent que le système est capable de fournir au médecin la meilleure action thérapeutique, et de fournir une simple mémorisation en opérant l'arbre de jonction cela en agrandissant sa base de cas.

## 5. Etude comparative

Le tableau (Tab.5.2) montre le taux de similarité des différents systèmes appliqués CBR/BN, ainsi que le nombre des nœuds utilisés. Nous remarquons que l'application de l'algorithme Pearl pour les systèmes de [DJE 07], [DJE 10] et [DJE 12c] a donné des résultats encourageants. Ces résultats montrent que le système est capable de fournir au médecin la meilleure action de diagnostic grâce à l'algorithme Pearl. Nous remarquons aussi que l'algorithme JLO dans [DJE 06] a donné des résultats encourageants, puisque sur une base de 5 cas nous avons obtenus un taux de similarité de 87.88%. Les résultats obtenus montrent que le l'algorithme JLO a donné des résultats satisfaisants. Ces résultats confirme que JLO est plus efficace en terme de complexité de structure (et par conséquent en terme de complexité de temps de calcul) pour la résolution de ce problème médical. Le système DIAPH [MER 05] a utilisé le RàPC comme le seul ou la technologie principale. Ce système est dédié au diagnostic des pathologies hépatiques utilisant une base de 55 cas. Les cas sont représentés par un ensemble symptomatologie cliniques et des caractéristiques extraites de l'image scanographiques du foie. Dans la phase de remémoration, deux mesures de similarités Block City et Euclidienne ont été appliqué produisant un taux de similarité de 98%. Le système PROBIS [MAL 96] intègre les réseaux de neurones dans le RàPC. Il contient une mémoire à deux niveaux de hiérarchie : le bas niveau qui comprend une mémoire plate des cas organisés en zones où chaque zone contient des cas similaires ; le haut niveau contient des prototypes, chacun représentant un groupe de cas du bas niveau. Il a permis de combiner les avantages de l'utilisation d'une mémoire hiérarchique (efficacité de remémoration avec un taux similarité de 86.11%) sans perdre l'avantage de l'utilisation d'une mémoire plate (précision de réponse). Le système CRN [NOU 10] est un système RàPC intégrant le réseau de recherche de cas nommé CRN « Case Retrieval Nets ». Cette structure CRN est une structure dynamique, facilitera ainsi la recherche des cas en entrée.

A travers cette étude comparative, nous signalons que l'intégration du réseau bayésien dans le paradigme RàPC a permis d'améliorer les performances de ce dernier.

| Modèles | Le nombre de cas | Le nombre de variables (nœuds) | Le taux de similarité |
|---|---|---|---|
| PROBIS [MAL 96] | 22 | - | 86.11% |
| DIAPH [MER 05] | 55 | 22 | 98% |
| Algorithme JLO [DJE 06] | 5 | 21 | 87.88% |
| Algorithme Pearl [DJE 07] | 25 | 56 | 79.68% |
| CRN [NOU 10] | 63 | 20 | 73.87% |
| Djebbar et al. [DJE 10] | 6 | 25 | 54.76% |
| Algorithme Extended Pearl [DJE 12c] | 19 | 42 | 60% |

Tab.5.2. Comparaison des résultats

Le tableau 5.3 montre la complexité de la modélisation de la phase de remémoration par l'algorithme Pearl [DJE 10][DJE 12a][DJE 12b] et l'algorithme JLO [DJE 06]. Nous remarquons que l'algorithme Pearl a une complexité encourageante. D'autre part l'arbre de jonction permet d'avoir un gain de temps important par rapport à la structure du réseau bayésien (la remémoration et la mémorisation du cas dans l'arbre de jonction se font dans un sous arbre où ses cliques contiennent au moins une variable du nouveau cas), ce qui permet d'améliorer les phases de remémoration et de mémorisation. Malgré que l'algorithme Pearl est spécifique aux arbres contrairement à JLO qui est plus générale mais la complexité en temps de calcul sera exponentielle dans les deux cas. Cette complexité est en fonction de la structure du graphe relativement à la fois au nombre de variables (nœuds).

| | Phase de remémoration | |
|---|---|---|
| | **Algorithme Pearl** | **Algorithme JLO** |
| Complexité | $n*2^m$ = $43*2^4$ =688 | $n*2^c$ = $21*2^3$ = 168 |

Tab.5.3. Calcul des complexités

Les mesures utilisées dans le tableau (Tab.5.3) représentent la complexité de la modélisation de la phase de remémoration afin d'évaluer les algorithmes utilisés tel que:

n : le nombre de nœuds du réseau.

m : le nombre maximum de parents (le degré).

c : la taille maximale de la clique de l'arbre.

Signalons que la complexité de l'algorithme Pearl est exponentielle et elle dépend du nombre des nœuds parents du graphe. Si le degré du graphe est constant, la complexité de l'algorithme Pearl prend un temps linéaire. Ainsi que l'algorithme JLO, sa complexité est exponentielle et elle dépend de la taille de la plus grande clique, toutefois elle devra linéaire si les tailles de cliques sont constantes.

## 6. Outil graphique du logiciel

Cette section sera consacrée à l'outil graphique permettant de manipuler et d'exécuter graphiquement toutes les méthodes implémentées. Cette application a été conçue sous Visuel C++ afin de définir l'ensemble de fonctionnalités et leurs interactions avec l'utilisateur, mettant en évidence un schéma d'enchainements de fenêtres simple et convivial.

Le menu principal de cette application est représenté par la figure (Fig.5.9). Nous expliquons le mode de fonctionnement de cet outil en décrivant ses diverses options.

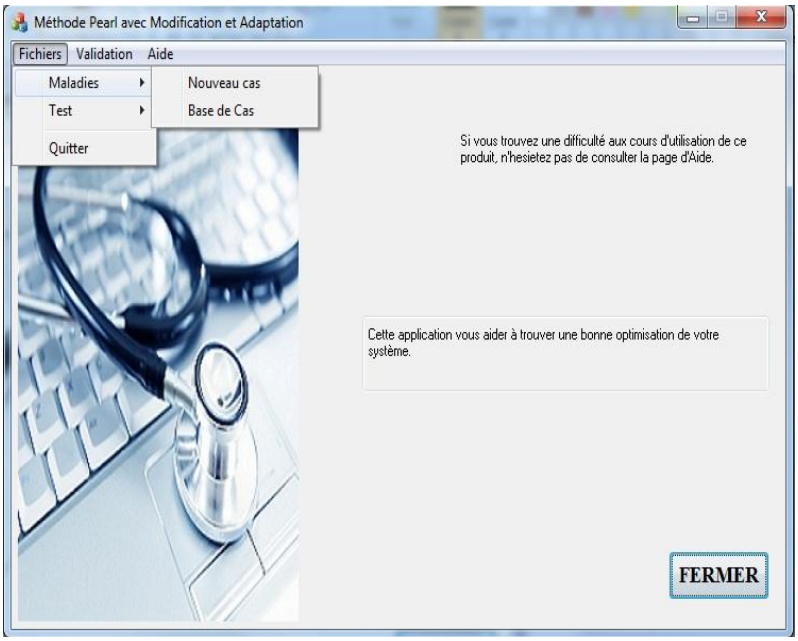



Fig.5.9. Menu principal

### 6.1. La structure du réseau

Le cas est décrit en termes de dimensions ou descripteurs (attributs), chaque descripteur peut prendre une probabilité qui a une valeur discrète. Ces attributs sont répartis en quatre niveaux qui sont élaborés dans le chapitre 4 (section 4.2). Chaque attribut représente soit un maître symptôme soit un non maître symptôme. Un exemple d'un graphe associé au réseau bayésien est décrit dans la figure (Fig.5.10).

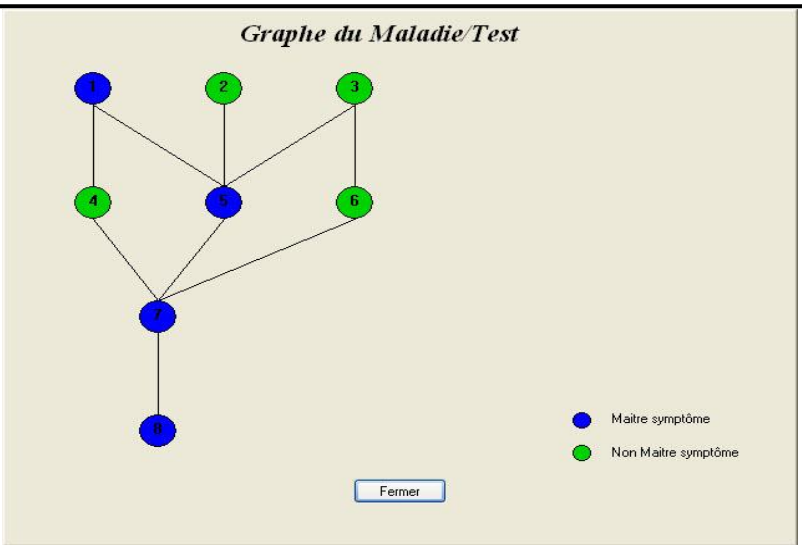



Fig.5.10. Exemple d'un graphe associé au réseau bayésien

Le tableau ci-dessous (Tab.5.4) illustre les différents nœuds de notre réseau bayésien.

| Id | Nœud | Id | Nœud | Id | Nœud | Id | Nœud |
|---|---|---|---|---|---|---|---|
| 01 | Découverte Fortuite | 12 | Sérologie | 23 | Taille TDM | 33 | Foie polykystique |
| 02 | Douleur | 13 | Hypoglycémie | 24 | Biologie génétique | 34 | Angiome |
| 03 | Ictère | 14 | Om+ov | 25 | Alpha foeto protéine | 35 | Adénome |
| 04 | Hpm | 15 | Dysphi | 26 | Kyste biliaire | 36 | Cfl carcinome |
| 05 | Alcool | 16 | Hépatomégalie | 27 | Kyste hydatique | 37 | Abcès hépatique |
| 06 | Ostroprostatie | 17 | Aghg | 28 | Cirrhose | 38 | Foie cardiaque |
| 07 | Age | 18 | Hépatite B/C | 29 | Glycogénose | 39 | Hyperbili antibinemie |
| 08 | Antécédent | 19 | Bilirubine | 30 | Métastase hépatique | 40 | Stéatose Hépatique |
| 09 | ACE | 20 | phosphatase | 31 | Amylase hépatique | 41 | Glycogénose hépatique |
| 10 | Insuffisance rénale | 21 | Homogénéité Echo | 32 | Cystadéno carcinome | 42 | Localisation hépatique des hémopathies |
| 11 | Infection | 22 | Homogénéité TDM | 33 | Foie polykystique | 43 | CHC Carcinome |

Tab.5.4. Les différents nœuds du réseau bayésien

### 6.2. La base de cas

Nous avons crée une base de cas contenant 19 cas avec 11 attributs pour chaque cas. La figure (Fig.5.11) donne un aperçu de la base de cas.

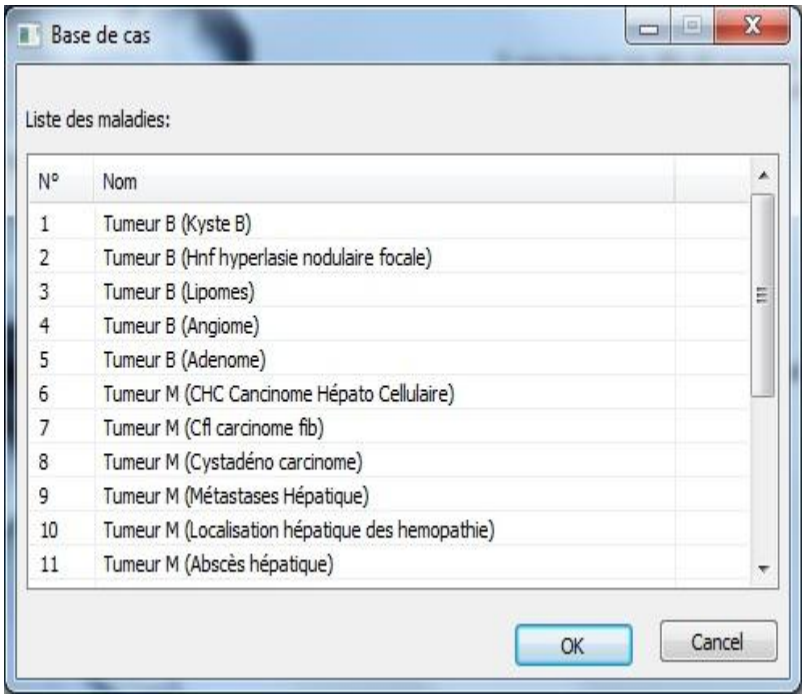



Fig.5.11. La base de cas

### 6.3. Ajout d'un nouveau cas

La fenêtre (Fig.5.12) permet de spécifier la structure initiale du réseau bayésien en donnant la possibilité à l'utilisateur d'ajouter une nouvelle maladie (nouveau cas), ou de supprimer des nœuds et des liens. Cette opération consiste à préciser la position du nœud à ajouter, en lui donnant un nom, une probabilité, un niveau, le type et définir ses parents parmi l'ensemble des nœuds existants.

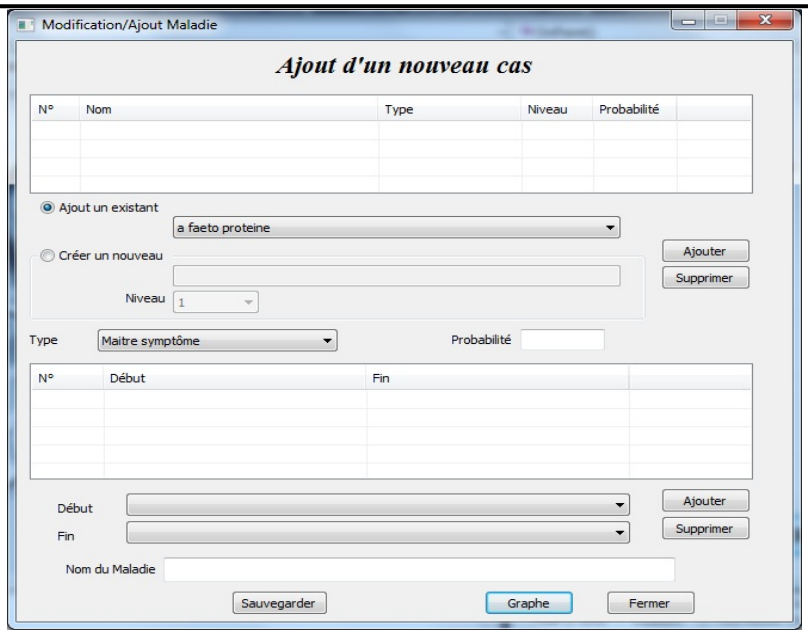



Fig.5.12. Possibilité d'ajout d'un nouveau cas

## 6.4. Autre graphique

Cette option (Fig.5.13) permet de spécifier la comparaison de deux algorithmes dans la phase de remémoration et la phase d'adaptation.

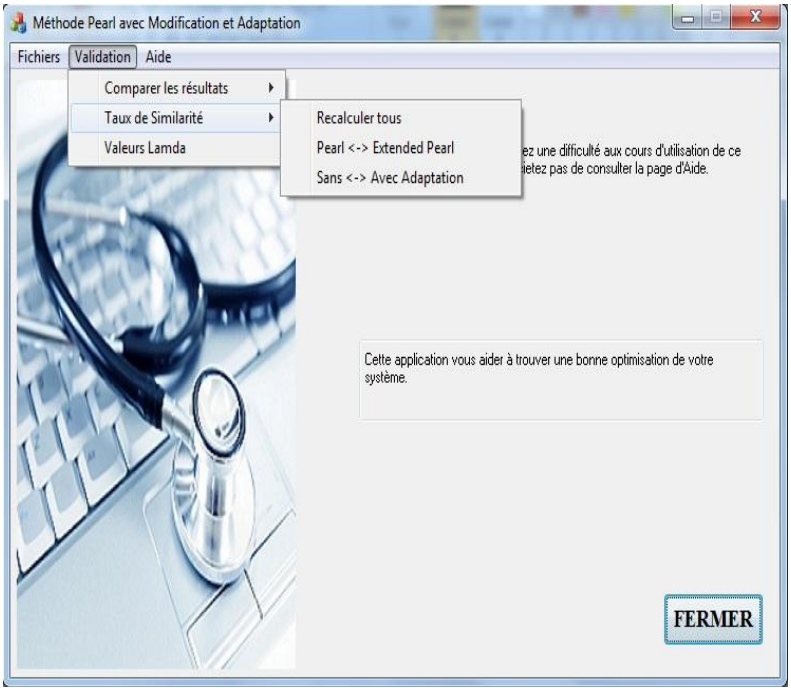



Fig.5.13. La phase de la remémoration et de l'adaptation

Les valeurs de lamda de la figure (Fig.5.14) représentent chacune le poids de chaque niveau du réseau bayésien. Ces valeurs sont déterminées à travers des tests pratiques.

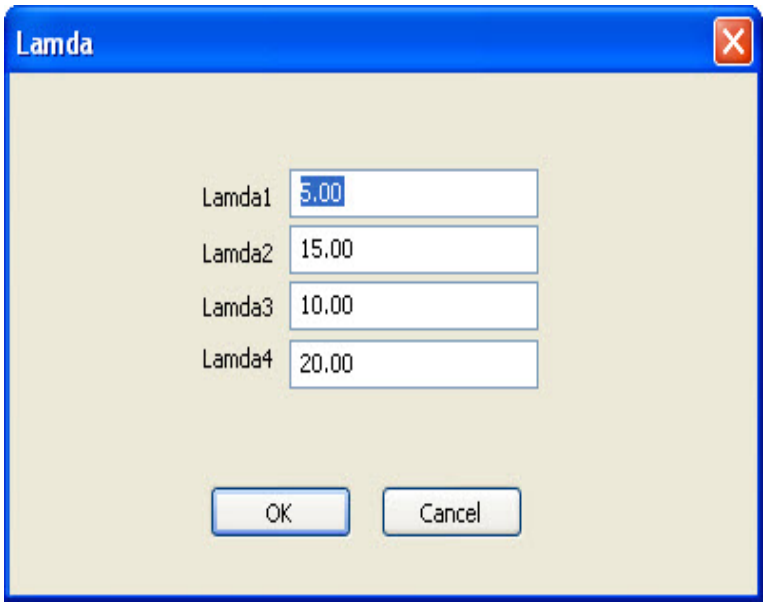



Fig.5.14. Les lamdas associés aux niveaux du réseau bayésien

La figure (Fig.5.15) montre un exemple d'un cas remémoré en utilisant l'algorithme Pearl.

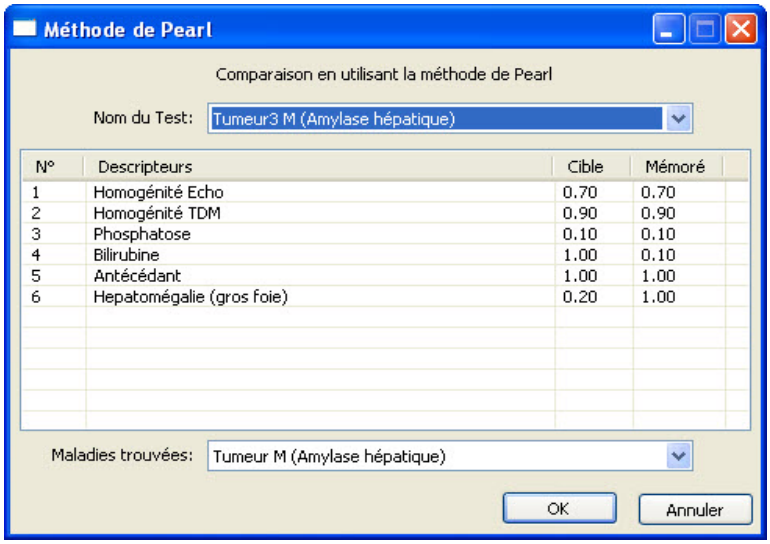



Fig.5.15. Exemple d'un cas remémoré

Notre application peut détecter qu'il n'existe pas le cas similaire au cas cible (Fig.5.16).

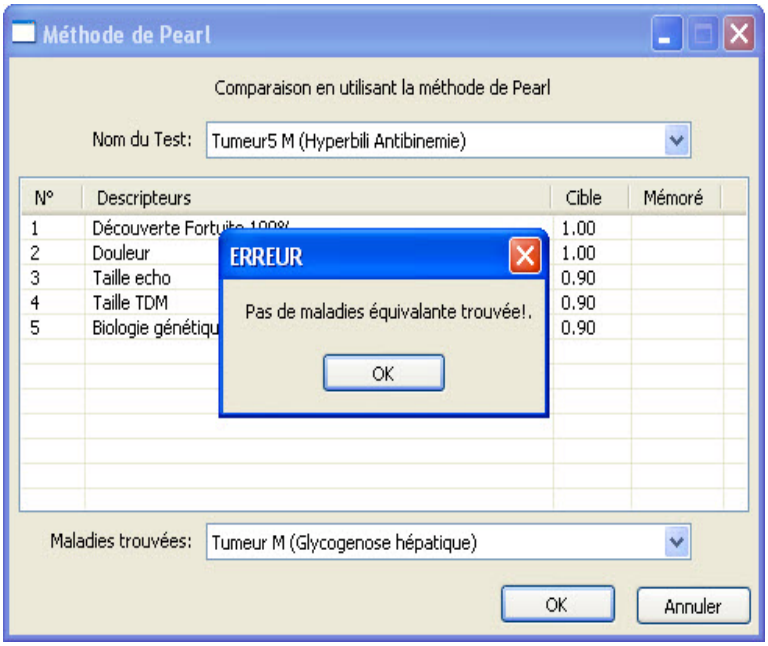



Fig.5.16. Détection d'un nouveau cas

## 7. Conclusion

Nous avons présenté dans ce chapitre l'implémentation des deux phases du cycle RàPC : la remémoration et l'adaptation. Pour la phase de remémoration, nous avons implémenté les deux algorithmes Pearl et Extended Pearl et nous avons établis une étude comparative avec d'autres modèles existants. Les résultats des tests d'évaluations ont montré que l'algorithme Extented Pearl proposé a fourni des résultats satisfaisants. Cet algorithme permet de calculer les probabilités marginales et assure un diagnostic efficace et facile pour les modèles graphiques sous forme de poly-arbres. Cet algorithme est appliqué sur le diagnostic des maladies les plus fréquentes du foie afin d'aider le médecin dans son processus de diagnostic. Pour la phase d'adaptation, nous avons abouti à des résultats encourageants (acceptables), cela signifie que pendant la phase de remémoration, qui prend en tenant compte de l'effort d'adaptation, peut ne pas sélectionner les cas les plus facilement adaptables.

Nous avons validé nos algorithmes sur un nombre de cas limité afin de prouver la faisabilité de notre proposition. Il reste à déployer l'application en grandeur nature, ce qui est une perspective immédiate de notre travail.

# Conclusion et perspectives

Nous résumons à présent les travaux de recherches présentés dans cette thèse sur l'intégration des réseaux bayésiens dans l'approche de raisonnement à partir de cas appliqué au diagnostic médical. Nous avons proposé une méthode de remémoration et d'adaptation à base de réseau bayésien.

Ce travail s'intègre dans le cadre de la modélisation des connaissances appliquée à l'aide au diagnostic des pathologies hépatiques. Il consiste à présenter deux phases d'un système RàPC d'aide aux médecins radiologues. La méthodologie de résolutions de problèmes adoptés s'appuie sur le paradigme RàPC fondé sur un modèle bayésien pour la modélisation des connaissances de la base de cas. La structure de ce modèle est un ensemble de connaissances utiles pour le diagnostic. Ces connaissances sont incertaines et elles sont réparties en quatre niveaux. Chaque niveau est décrit par un ensemble de variables. Nous avons organisé nos travaux en deux parties distinctes.

La première partie introduit les principes fondamentaux de l'approche RàPC. Cette introduction est suivie par le domaine d'étude qui est le diagnostic médical et un état de l'art sur les systèmes de diagnostic médical par le RàPC. La première partie est organisée à travers trois premiers chapitres. De plus, cette partie présente nos choix et nos démarches concernant la formalisation des deux phases de remémoration et d'adaptation à base de réseau bayésien pour l'aide au diagnostic des pathologies hépatiques.

La deuxième partie synthétise l'essentiel de notre contribution. En effet, cette contribution apparait au cours du quatrième et cinquième chapitre.

Les contributions de ce travail se résument dans les points suivants :

**Modélisation de la base de cas par un réseau bayésien**

Notre première contribution consiste à choisir un modèle structurel de mémoire de cas. L'idée principale consiste à modéliser la base de cas par un réseau bayésien. Les réseaux bayésiens sont d'excellents outils de modélisation de l'incertain grâce à leur représentation graphique claire et aux lois de probabilités conditionnelles définies sur ce graphe. Le réseau proposé permet une représentation de connaissances qualitatives causales et de connaissances quantitatives exprimant l'incertitude. Il est constitué de quatre niveaux, niveau clinique, niveau biologique, niveau imagerie médicale et niveau diagnostic et thérapie. Chaque niveau est constitué d'un ensemble d'attributs qui correspond chacun à un nœud du réseau. Les arcs

décrivent les relations entre ces attributs comme étant des probabilités conditionnelles des attributs dans le cas.

**Modélisation bayésienne de la phase de remémoration**

Le deuxième axe de contribution s'intéresse particulièrement à l'étape remémoration ainsi au choix de mesure de similarité et d'algorithme de recherche de cas similaires. Une modélisation de la remémoration en raisonnement à partir de cas repose sur un modèle bayésien via un modèle log linéaire pour la mesure de similarité. Cette modélisation consiste à modéliser les cas par une représentation hiérarchique. Nous avons modélisé la mesure de similarité par un modèle log linéaire qui consiste à exprimer des logarithmes des probabilités. Cette modélisation a l'avantage d'être hiérarchique et elle permet de maîtriser la complexité du modèle. Ce qui rend la remémoration plus efficace. La phase de remémoration consiste à sélectionner le cas le plus similaire du modèle Log-linéaire afin d'améliorer cette phase, ce qui fournit un diagnostic plus précis.

**Modélisation bayésienne de la phase d'adaptation**

La troisième contribution est issue à travers la phase de remémoration. Nous avons proposé un algorithme d'adaptation qui s'appuie sur une mesure d'adaptation. Cette mesure est basée sur deux paramètres essentiels qui sont le poids du niveau et le nombre de maîtres symptômes (ces deux paramètres sont primordiaux dans la détermination des étapes d'adaptations). Cet algorithme unifié l'étape de la remémoration et l'adaptation. Il se base sur la dépendance entre la description du problème du cas cible et la description du problème du cas source remémoré par la mesure de remémoration et la mesure d'adaptation proposée. La mesure d'adaptation proposée a pour objectif de choisir, parmi les cas remémorés le cas le plus facilement adaptable. Nous avons présenté trois exemples de diagnostic des pathologies hépatiques illustrant l'algorithme proposé.

Notre approche a été implémentée et testée sur une base de cas définie précédemment. Le but de cette approche est de fournir un diagnostic le plus précis possible et avec la plus grande certitude. Nous conclurons que la structure bayésienne de la base de cas facilite la recherche des cas similaires au cas d'entrée en sélectionnant le cas le plus probable. Une telle structure facilite l'adaptation en adaptant la solution du problème similaire afin de résoudre le nouveau cas. A travers les résultats obtenus, nous avons conclu que les cas les plus similaires au cas cible ne sont pas forcément ceux choisie pour la phase d'adaptation.

Parmi les perspectives envisageables à partir du travail réalisé, nous citons :

- Proposition de plusieurs modèles de structures du réseau bayésien de plusieurs experts pour le diagnostic des maladies du foie. Puis combiner ces modèles en une seule structure qui maximise la vraisemblance du domaine. Une fois la structure du réseau bayésien établie, il faut alors la paramétrer par des données empiriques. Pour avoir un modèle bayésien complet, il suffit de définir des tables de probabilités conditionnelles à chaque nœud du réseau.

- Notre travail représente une collaboration avec une équipe de médecins radiologues afin d'avoir des valeurs numériques qui sont les paramètres (les probabilités) du réseau bayésien. De ce fait, une étude statistique sur une population est indispensable afin de déterminer ces probabilités.

- On notera que le problème d'adaptation des paramètres du réseau est le problème clé dans notre système RàPC. Cette adaptation passerait par l'utilisation d'informations qualitatives multiples, fournies en général par le médecin, pour adapter les paramètres du réseau et permettre ainsi la production d'un diagnostic plus fiable. Ce type de problème n'a actuellement pas encore de solution et reste largement ouvert. La définition d'une ontologie de qualification d'un diagnostic serait donc nécessaire, et permettrait ainsi d'aider à inférer les modifications des paramètres nécessaires afin de mieux cibler les spécificités de chaque cas.

- La formalisation de l'étape de maintenance de la base de cas par les réseaux bayésiens: notre orientation principale des travaux futurs est certainement le développement de l'étape de maintenance en appliquant les techniques d'apprentissage des connaissances afin de permettre au système RàPC proposé d'effectuer les opérations de maintenance.

# Annexe A

# La probabilité conditionnelle et le théorème de Bayes

Le premier élément du langage de la probabilité est l'événement, un exemple classique d'événement est le lancement d'une pièce de monnaie en l'air.

Sachant l'événement *e,* on appelle *e* l'événement que *e* n'a pas lieu. La probabilité d'un événement *e* représente la fréquence que l'événement *e* a lieu. Il existe une autre représentation de la probabilité: la probabilité d'un événement *e* représente le degré de certitude d'une personne que l'événement *e* a lieu dans un seul essai. Si une personne assigne la valeur 1 à l'événement *e*, alors il est sur que *e* aie lieu. Par contre s'il assigne la valeur 0 à l'événement *e*, il croit que *e* n'a jamais lieu. Et s'il assigne la valeur entre 0 et 1 à l'événement *e*, peut-être il ne sait pas si *e* a lieu. On écrit $p(e)$ pour montrer la probabilité de l'événement *e*. Alors on a la règle de somme:

$$p(e) + p(\bar{e}) = 1 \qquad (1)$$

Soient 2 événements *a* et *b* on suppose que *b* s'est produit, s'il existe un lien entre *a* et *b* cette information va modifier la probabilité de *a*. Alors on a le concept de la probabilité conditionnelle:

$$p(a|b) = \frac{p(a,b)}{p(b)} \qquad (2)$$

Dans la formule (2) $p(a,b)$ représente la probabilité que tous les deux événements *a* et *b* ont lieu, $p(a|b)$ est la probabilité de *a* sachant *b*. À partir de la formule (2) on peut tirer le théorème de Bayes suivant:

$$p(a|b) = \frac{p(b|a).p(a)}{p(b)} \quad (3)$$

On dit que 2 événements *a* et *b* sont indépendants si et seulement si: $p(a,b)=p(a).p(b)$

En utilisant la définition de la probabilité conditionnelle on a:

$$p(a|b) = p(a) \text{ et } p(b|a) = p(b) \quad (4)$$

La formule au-dessus veut dire que la connaissance de *b* n'apporte rien sur celle de *a* et réciproquement.

Pour représenter les événements d'un essai on utilise les variables aléatoires. Une variable aléatoire accepte une valeur d'un ensemble complet, mutuellement exclusif des états, chaque état est correspondant à un ou plusieurs événements. Une variable peut être discrète ou continuelle.

Sachant que la variable Y peut accepter un des états $y_1..y_n$, on a le théorème des probabilités totales:

$$p(X) = \sum_Y p(X|y).p(y) \quad (5)$$

À partir de la formule (3) et la formule (5), finalement on a le théorème de Bayes complet suivant:

$$p(Y|X) = \frac{p(Y).p(X|Y)}{\sum_Y p(X|y).p(y)} \quad (6)$$

# Annexe B

# Les termes médicaux

| | |
|---|---|
| **Abcès** | Lésion arrondie remplie de liquide fréquemment volumineux. Les abcès de candidose sont multiples. |
| **Adénome** | C'est une tumeur rare à forte prédominance féminine, favorisée par la prise d'oestroprogestatifs, le plus souvent unique. C'est une prolifération bénigne d'hépatocytes normaux richement vascularisée pouvant contenir de la graisse et survenant sur un foie sain. |
| **Angiome hépatique** | C'est une tumeur bénigne très fréquente constituée de petites cavités remplies de sang. Il est fréquemment sous-capsulaire et multiple dans 20% des cas et lorsqu'il est inférieur à 4 cm totalement asymptomatique. Les angiomes de grande taille (> 4 cm) peuvent être le siège de remaniements fibreux et nécrotiques et devenir symptomatiques. |
| **Antécédent** | Événement pathologique, personnel ou familial, qui, chez un sujet, a précédé sa maladie actuelle. |
| **Atrophique** | Un foie de petite taille ou atrophique. IL se rencontre le plus souvent dans les cirrhoses évoluées. |
| **Bilirubine** | La bilirubine dérive du catabolisme de l'hème, essentiellement de l'hémoglobine. Dans le plasma, elle est transportée, non conjuguée et insoluble, liée à l'albumine. Elle est captée par l'hépatocyte, conjuguée et excrétée dans la bile. La bilirubine conjuguée est soluble dans l'eau. En cas de lésion hépatocytaire ou d'obstacle à l'écoulement biliaire, la bilirubine conjuguée reflue dans le plasma. |
| **Carcinome hépatocellulaire (CHC)** | C'est la tumeur maligne primitive du foie la plus fréquente, survenant dans 80% des cas sur un foie de cirrhose d'origine divers (alcool, hépatites B, C, hémochromatose). Il peut être unique ou multifocal, nodulaire ou infiltrant. En échographie, son aspect est variable : Nodule hypoéchogène homogène (si inf à 3 cm) ou hyperéchogène ou hétérogène à centre nécrotique, parfois cerné d'une bande hypoéchogène correspondant à une capsule fibreuse. La présence d'une thrombose portale ou sus-hépatique associée est un argument supplémentaire pour le diagnostic. |

| | |
|---|---|
| **Cirrhose** | Résulte principalement d'épisodes répétés d'hépatite alcoolique ; elle a pour conséquences l'hypertension portale et l'insuffisance hépatocellulaire; elle est irréversible. |
| **Dyspnée** | La dyspnée (en latin *dyspnoea*, en grec *dyspnoia* de *dyspnoos* - court d'haleine) est une difficulté respiratoire. |
| **Foie** | Le plus gros organe interne chez les vertébrés, pesant environ 1,5 kg. De couleur rouge sombre, il est situé dans le quart supérieur droit de l'abdomen. Le foie a un débit sanguin moyen de 1,4 litre par minute ; il contient en permanence environ 10 p. 100 du sang total de l'organisme. Il est également traversé par le sang du pancréas et de la rate. Les cellules hépatiques aident le sang à assimiler les substances nutritives, et à excréter les déchets et les toxines, ainsi que des substances comme les stéroïdes, les œstrogènes et d'autres hormones. |
| **Foie cardiaque** | Le foie cardiaque congestif est défini par l'ensemble des manifestations hépatiques secondaires à une élévation de la pression veineuse centrale. |
| **Hépatique** | Artère qui amène au foie le sang oxygéné provenant du cœur. |
| **Hépatite** | L'hépatite (du grec *hépar*: foie) désigne toute inflammation aiguë ou chronique du foie Les formes les plus connues étant les formes virales (notées de A à E) et alcoolique. Mais l'hépatite peut aussi être due à certains médicaments, un trouble du système immunitaire de l'organisme .L'hépatite est dite aiguë lors du contact de l'organisme avec le virus et chronique lorsqu'elle persiste au-delà de 6 mois après le début de l'infection. L'hépatite peut évoluer ou non vers une forme grave ou fulminante, une cirrhose ou un cancer. |
| **Hépatite alcoolique** | Survient en cas d'intoxication importante. Elle peut être symptomatique ou évoluer à bas bruit. Le foie est le plus souvent déjà cirrhotique. Dans ses formes sévères, elle peut être mortelle (dans 1 cas sur 2 environ). |
| **Hépatite toxique** | L'hépatite toxique est une inflammation du foie causée par des produits chimiques. De nombreux produits chimiques inhalés ou ingérés de façon délibérée ou non peuvent avoir des effets toxiques sur le foie. Parmi ces produits chimiques figurent les médicaments, les solvants industriels et les polluants. |
| **Hépatites virales** | Regroupent les infections provoquées par des virus se développant aux dépens du tissu hépatique. Les virus, une fois inoculés à l'organisme, infectent alors préférentiellement les cellules du foie aussi appelées hépatocytes. |

| | |
|---|---|
| **Hépatomégalie** | Un foie de grande taille ou hépatomégalie. IL se voit dans diverses pathologies (cirrhose, hépatite aiguë, foie tumoral diffus, foie cardiaque…). |
| **Hyperplasie nodulaire focale (HNF)** | C'est une tumeur bénigne touchant préférentiellement la femme, secondaire à une malformation artérielle aboutissant à une hyperplasie du parenchyme hépatique. |
| **Hypertension portale** | L'hypertension portale est définie soit par une augmentation de la pression portale au-delà de 15 mmHg, soit par une élévation du gradient de pression porto-cave au-delà de 5 mmHg. |
| **Ictère** | Un ictère ou jaunisse correspond à la coloration *jaune* des téguments (peau et muqueuses : on parle d'ictère cutanéo-muqueux) due à l'accumulation de bilirubine, qui peut être libre ou conjuguée. |
| **Lésion kystique** | Le kyste biliaire est le plus fréquent, le plus souvent asymptomatique de découverte fortuite. En échographie, le kyste biliaire se présente comme une lésion anéchogène (vide d'échos) à paroi fine, sans cloison, présentant un renforcement postérieur des échos. |
| **Métastases hépatiques** | Souvent multiples, les métastases hépatiques les plus fréquentes sont celles de cancer colorectal, pulmonaire, mammaire ou pancréatique. Toutefois, toute tumeur maligne évoluée peut s'accompagner de métastases hépatiques. |
| **Phosphatase** | Une phosphatase est une enzyme dont la fonction est d'enlever un groupe phosphate d'une molécule simple ou d'une macromolécule biologique, par hydrolyse. En biotechnologie, elle peut agir sur l'ADN et sur l'ARN afin d'empêcher leur circularisassions. Elle est utilisée après l'action d'une enzyme de restriction et avant l'addition de l'ADN étranger d'intérêt. |
| **Sérologie** | La sérologie est littéralement l'étude du sérum, qui est une partie du plasma sanguin. En pratique, c'est la recherche d'anticorps dans ce sérum, dirigés contre des microbes ou, dans le cas des maladies auto-immunes, contre les propres constituants de l'organisme. |

# Bibliographie

**[AAM 94]** Aamodt A. et Plaza E., Case Based Reasoning: Foundational issues, methodological variations, and system approaches. AI communications, IOS Press, Vol 7(1), pp. 39-59, 1994.

**[AAM 98]** Aamodt A. et Langseth H., Integrating Bayesian Networks into Knowledge-Intensive CBR. In: AAAI Workshop on Case-Based Reasoning Integrations, 1998. Technical Report WS-98-15. AAAI Press, Menlo Park, ISBN 1-57735-068-5, pp. 1-6, 1998.

**[ABE 96]** Abecassis P. et Batifoulier P.**,** Comment penser l'incertitude médicale à l'aide des probabilités ?. 1996.

**[AFO 04]** Afouba N., Kerbrat S. et Labarang Z., Rapport du projet d'intelligence artificielle : Le raisonnement à partir des cas : Définitions et principes de fonctionnement, Novembre 2004.

**[AGN 98]** Agnieszka O., Malek J., Druzdzel J. et Wasyluk H., A Probabilistic Causal Model for Diagnostic of Liver Disorders. Intelligent Information Systems VII, Proceedings of the Workshop held in Malbork, Poland, pp. 379-387, 1998.

**[AHA 96]** Aha D. et Chang L., Cooperative bayesian and case-based reasoning for solving multiagent planning tasks. Technical Report AIC-96-005, Navy Center for Applied Research in Artificial Intelligence, 1996.

**[ALE 10]** Aleksovska S. L. et 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., Idexation guidée par les connaissances en Imagerie médicale. Thèse de doctorat, Ecole Doctorale-sicma, L'Université européenne de Bretagne, Janvier 2012.

**[ARM 09]** Armaghan N., Contribution à un système de retour de l'expérience basé sur le raisonnement à partir de cas conversationnel : application à la gestion des pannes de machines industrielles. Ecole doctorale Science et Ingénieur (RP2P), Université de Nancy, Mai 2009.

**[BAL 03]** Balaa Z. et al., FM-Ultranet: A decision support system using case-based reasoning applied to ultrasonography. In: McGinty, L. (Ed.): Workshop Proceedings of the International Conference on Case-Based Reasoning, pp. 37-44, 2003.

**[BAR 88]** Bareiss E. R., PROTOS: A Unified Approach to Concept Representation, Classification and Learning. PhD Thesis, University of Texas, 1988.

**[BEG 09]** Begum S., Mobyen Uddin A., et Funk P., Case-Based Systems in the Health Sciences: A Case Study in the Field of Stress Management. WSEAS Transactions on Systems, Vol 8, nr 1109-2777, pp. 344-354, March, 2009.

**[BEL 02]** Bellot D., Fusion de données avec des réseaux bayésiens pour la modélisation des systèmes dynamiques et son application en télémédecine. Thèse de doctorat à l'université Henri Poincaré, Nancy, 22 Novembre 2002.

**[BIC 94]** Bichindaritz I., Apprentissage de concepts dans une mémoire dynamique : Raisonnement à Partir de Cas adaptable à la tâche cognitive. Thèse de doctorat, Université René Descartes, 1994.

**[BIC 03]** Bichindaritz I., Solving safety implications in a case based decision-support system in medicine, 2003.

**[BLU 06]** Blunsom P., Kocik K. et Currran J.R., Question classification with Log-Linear Models. In SIGIR,06, Seattle, Washington, USA, 2006.

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

**[BOZ 98]** Bozec C., Jaulent C.M., Erie Z. et Partrice D., Accès Internet à IDEM images et diagnostics par l'exemple en médecine. Santé et Réseaux Informatiques, Vol 10, 1998.

**[BRA 93]** Bradburn C. et Zeleznikow J., The application of case-based reasoning to the tasks of health care planning. In: Wess, S., Althoff, K.-D., Richter, M.M. (Eds.): Topics in case-based reasoning. Proceedings of the EuropeanWorkshop on Case-Based Reasoning, EWCBR-93. Lecture Notes in Artificial Intelligence Springer,Vol 837, pp. 365–378, Berlin Heidelberg New York, 1993.

**[BRE 00]** Bresson B. et Lieber J., Raisonnement à partir de cas pour l'aide au traitement du cancer du sein. In : Actes des journées ingénierie des connaissances, pp. 189-196, 2000.

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

**[BRU 09]** Bruland T., Aamodt A. et Langseth H., Architectures Integrating Case-Based Reasoning and Bayesian Networks for Clinical Decision Support, pp. 1-10, Norway, 2009.

**[BUI 04]** Buist E., Les éléments fondamentaux du raisonnement à base de cas Les éléments fondamentaux du raisonnement `a base de cas. pp. 1-46, Février 2004. http://www.ericbuist.com/me/travaux/cbr.pdf

**[CHR 97]** Christensen R., Log-Linear Models and Logistic Regression. Springer, pp 1-7, 1997.

**[COO 90]** Cooper G., Computational complexity of probabilistic inference using Bayesian belief networks. In Artificial Intelligence, Vol 42(2), pp. 393-405, 1990.

**[COR 03]** Corset F., Optimisation de la maintenance à partir de réseaux bayésiens et fiabilité dans un contexte doublement censuré. PhD thesis, Université Joseph Fourier, 2003.

**[DAQ 04]** D'Aquin M., Brachais S., Lieber J., et Amedeo N., Vers une acquisition automatique de connaissances d'adaptation par examen de la base de cas: une approche fondée sur des techniques d'extraction de connaissances dans des bases de données. In Kanawati, R., Salotti, S., & Zehraoui, F. (Eds.), Actes du douzième atelier raisonnement à partir de cas, RàPC'04, pp. 41-52, 2004.

**[DEC 07]** Déchelotte D., Traduction automatique de la parole par méthodes statistiques. Thèse de doctorat, Université Paris Sud 11, Faculté des sciences d'Orsaypp, pp. 16-23, 2007.

**[DEL 05]** Delcroix V., Maalej M. A. et Piechowiak S., Réseaux bayésiens versus d'autres modèles probabilistes pour le diagnostic multiple de systèmes complexes. Revue des Nouvelles Technologies de l'Information (RNTI-E), 5, pp. 65-68, 2005.

**[DJE 04]** Djebbar A., Grainia S. et Merouani H. F., Un système de reconnaissance des images scanographiques appliqué au diagnostique des pathologies hépatiques. JETIM'04: Journées d'études Algéro-Francaises en imageries médicale, pp.131-136, Blida, 2004.

**[DJE 06]** Djebbar A. et 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 07]** Djebbar A., Toumiette A. et Mansri W., Conception et Implémentation d'un Système D'aide Au Diagnostic Des Pathologies Hépatiques à base de Réseau Bayésien. Rapport de recherche interne, SRF/LRI, 2007.

**[DJE 10]** Djebbar A., Refai A. et 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.

**[DJE 11a]** Djebbar A. et Merouani H.F., Une modélisation bayésienne de la remémoration pour l'aide au diagnostic. ICITeS'2011, April 10-12, Sousse, Tunisia, 2011, www.dline.info/content.htm.

**[DJE 11b]** Djebbar A. et Merouani H.F., Un modèle statistique de la remémoration d'un RàPC pour l'aide au diagnostic médical. JDLIO' 2011, 1ères Journées Doctorales du Laboratoire d'Informatique d'Oran, 31 Mai et 01 Juin 2011.

**[DJE 11c]** Djebbar A. et Merouani H.F., Une Remémoration Bayésienne d'un RàPC pour l'Aide au Diagnostic Médical. R2I' 2011, Rencontres sur la Recherche en Informatique $R^2I$ 201, Tizi-Ouzou, Algérie. 12-14 Juin 2011.

**[DJE 12a]** Djebbar A. et Merouani H.F., Retrieval of cases by using a Bayesian Network applied to the medical diagnosis. The Mediterranean Journal of Computers and Networks Medjcn, Vol 8(2), ISSN: 1744-2397, pp. 69-74, 2012.

**[DJE 12b]** Djebbar A. et Merouani H.F., Applying BN in CBR Adaptation-Guided Retrieval for medical diagnosis. International Journal of Hybrid Information Technology, Vol 5(2), ISSN: 1738-9968, pp. 41-56, 2012.

**[DJE 12c]** Djebbar A and Merouani H.F., Retrieval and adaptation in CBR through Bayesian Network for diagnosis of hepatic pathologies. International Journal of Hybrid Intelligent System, IOS press, Vol 9(3), ISSN: 1448-5869, pp. 123-134, 2012.

**[DJE 12d]** Djebbar A. et Merouani H.F., Optimizing retrieval phase in CBR through Pearl and JLO algorithms for medical diagnosis. International Journal of Advanced Intelligence Paradigms (IJAIP), Inderscience, Vol 4(3/4), ISSN: 1755-0386, 2012.

**[DJE 12e]** Djebbar A. et Merouani H. F., BN-CBR model for diagnosis of hepatic pathologies, Biomedical Engineering International Conference (BIOMEIC'12), ISSN: 2253-0886, October 10-11, Tlemcen, Algerie, 2012.

**[DON 10]** Dong D., Zhaohao S. et Feng G., PCOPM: A Probabilistic CBR Framework for Obesity Prescription Management. In Proceedings of ICIC (2), pp.91-99, 2010.

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

**[FUC 08]** Fuchs B., Raisonnement à Partir de Cas, In Renaud, J., Chabel Morello B, et Metta N, (éd), Retour et capitalisation d'expérience, outils et démarche. La Plaine Saint-Denis : AFNOR, pp. 184, 2008.

**[GEB 97]** Gebhardt F., Voß A., Gräther W. et Schmidt-Belz B., Reasoning With Complex Cases. Kluwer academic publishers, 1997.

**[GIE 98]** Gierl L., Bull M. et Schmidt R., CBR in Medicine. In Case-Based Reasoning Technology, pp. 273-298, 1998.

**[GOM 04]** Gomes P., Software design retrieval using Bayesian Networks and WordNet. Lecture Notes in Computer Science, pp. 184-197, 2004.

**[GRA 10]** Gravem A. M., Integrating Case-based and Bayesian Reasoning for Decision Support, Master of Science in Computer Science, Department of Computer and Information Science, Norwegian University of Science and Technology, June 2010.

**[HAD 97]** Haddad M., Adlassnig K.P. et Porenta G., Feasability analysis of a case-based reasoning system for automated detection of coronary heart disease from myocardial scintigrams. Artificial Intelligence in Medicine 9, pp. 61-78, 1997.

**[HAJ 04]** HajSaid A., Distances sémantiques pour la comparaison des connaissances objets dans le cadre du raisonnement à partir de cas. Mémoire de DEA Informatique Théorique et Applications, Université de le Havre, pp. 25-40, 2004.

**[HAM 89]** Hammond K.J., Adaptation of Cases. In: Proceedings of DARPA Workshop on Case- Based Reasoning, pp. 88–89, Florida, June 1989.

**[HAO 09]** Haouchine M. K., Remémoration guidée par l'adaptation et maintenance des systèmes de diagnostic industriel par l'approche du raisonnement à partir de cas. L'UFR des Sciences et Techniques de l'Université de Franche-Comté, Septembre 2009.

**[HAT 91]** Haton J. P., Bouzid N., Charpillet F., Haton M. C. B., Laasri H., Marquis P., Mondot T. et Napoli A., Le raisonnement en Intelligence Artificielle, pp. 8-9, InterEditions, Paris, 1991.

**[HEI 99]** Heinisch R. H., Weber R., Martin A. et Barcia R. M., Industrial Engineering and Cardiology Service of University Hospital. Florianopolis, SC, Brazil, 1999.

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

**[HEN 02]** Hennessy D., Buchanan B. et Rosenberg J., Bayesian Case Reconstruction. Lecture Notes in Computer Science, Springer, Heildelberg, pp.148–158, 2002.

**[JEN 90]** Jensen F., Lauritzen S. L. et Olesen K., Bayesian updating in causal probabilistic networks by local computations. Computational Statistics Quaterly, Vol 4, pp. 269-282, 1990.

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

**[JUR 98]** Jurisica I. et al., Case-based reasoning in IVF: Prediction and knowledge mining. In Artificial Intelligence in medicine, Vol 12(1), pp. 1-24, 1998.

**[KAH 93]** Kahn CE.Jr., Case-based selection of diagnostic imaging procedure. In: Leake DB(ed). Case-Based Reasoning: Papers from the 1993 Workshop. Menlo Park: AAAI Press, 1993.

**[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.

**[KEY 02]** Keysers D. et Paredes R., Comparaison of Log-linear Models and Weighted Dissimilarity Measure. Iberian Conference on Pattern Recognition and Image Analysis, pp. 370–377, 2002.

**[KIN 67]** King L. S., What is a diagnosis?. JAMA: The Journal of the American Medical Association, Vol 202(8), pp. 714-717, Novembre 1967.

**[KOL 88]** Kolodner J., Workshop on case-based Reasoning. Editor *DARPA 88*, Clearwater, Morgan Kaufmann, San Mateo, Florida, 1988.

**[KOL 93]** Kolodner J., Case-Based Reasoning. Morgan Kaufmann, San Mateo, Publishers, Inc, UCA, 1993.

**[KON 08]** Kong G., Xu D. L. et Yang J. B., Clinical decision support systems: a review on knowledge representation and inference under uncertainties. International Journal of Computational Intelligence Systems, Vol 1(2), pp. 159-167, 2008.

**[KOT 88]** Konton P., Reasoning about Evidence in Causal Explanations. In proceedings of the workshop on Case-Based reasoning (DARPA), pp. 260-270, 1998.

**[KOT 03]** Kotchi C. S., Delcroix V. et Piechowiak S., Etude de la performance des algorithmes d'inférence dans les réseaux bayésiens. 2003.

**[LAU 88]** Lauritzen S. L. et Spiegelhalter D., Local computations with probabilities on graphical structures and their application to expert systems. Journal of the Royal Statistical Society: Series B, Vol 50(2), pp. 157-224, 1988.

**[LEA 01]** Leake D. B., Smyth B., Yang Q. et Wilson D., Special Issue on Maintaining Case- Based Reasoning Systems. Computational Intelligence, Vol 17(2), 2001.

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

**[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 99]** Leray PH. et Gallinari P., Une architecture neuro-bayésienne pour le traitement spatio-temporel d'alarmes, Application au diagnostic dans le réseau téléphonique. In Journées Nationales sur les Modèles de Raisonnement, JNMR'99, pp. 134-145, 1999.

**[LER 02]** Leray P. et Olivier F., Réseaux Bayésiens pour la Classification : Méthodologie et Illustration dans le cadre du Diagnostic Médical. RIA, 15/2002, Réseaux Bayésien, pp. 1-25, 2002.

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

**[MAL 96]** Malek M., Un modèle hybride de mémoire pour le raisonnement à partir de cas. Thèse de doctorat, Université Joseph Fourier, 30 octobre 1996.

**[MER 05]** Merouani H.F., Djebbar A., Sayed S. et Dhouaibia F., DIAPH : Système à base de cas pour l'aide au diagnostic des pathologies hépatiques. Rapport de Recherche Interne, LRI/SFR, 2005.

**[MEY 09]** Meyer N., Vinzio S. et Goichot B., La statistique bayésienne : une approche des statistiques adaptée à la clinique. La revue de médecine interne, Vol 30(3), pp. 242-249, 2009.

**[MIL 07]** Mille A., Les Réseaux Bayésiens A la recherche de la vérité. Cours Cognition et connaissance, Université Claude Bernard Lyon 1, pp. 1-9, 2007.

**[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.

**[MIN 75]** Minsky M., A framework for representing knowledge. In the psychology of Computer Vision, Editeur P.H. (Ed.) Winston, pp. 211-279, New York, McGraw Hill, 1975.

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

**[NIL 04]** Nilson M. et Sollenborn M., Advancements and Trends in Medical Case Resoning: An Overview of Systems and System Development. 2004.

**[NOU 04]** Nouaouria N. et Laskri M. T., Un réseau de recherche de cas orienté adaptapilité. 12ème Atelier de Raisonnement à Partir de Cas, LIPN, Université Paris13, Villetaneuse, 2004.

**[NOU 10]** Nouaouria N. et Boukadoum M., Case Retrieval with Combined Adaptability and Similarity Criteria: Application to Case Retrieval Nets. I. Bichindaritz and S. Montani (Eds.): ICCBR 2010, LNAI 6176, Springer-Verlag, Berlin Heidelberg, pp. 242-256, 2010.

**[OCA 11]** Ocampo M., Maceiras M., Herrera S., Maurente C., Rodríguez D. et Sicilia M., Comparing Bayesian inference and case-based reasoning as support techniques in the diagnosis of Acute Bacterial Meningitis. Elsevier journal, pp. 1-12, 2011.

**[OPI 95]** Opiyo E.T.O., Case-based reasoning for expertise relocation in support of rural health workers in developing countries. In: Aamodt, A., Veloso, M. (Eds.): Casebased reasoning research and development. Proceedings of the International Conference on Case-Based Reasoning, ICCBR-95. Lecture Notes in Artificial Intelligence, Vol 1010, Springer, Berlin Heidelberg New York, pp. 77-87, 1995.

**[PAC 04]** Pacquet S., Raisonnement probabiliste. Intelligence artificielle II, IFT-17587, pp. 1-64, 2004.

**[PAL 04]** Pal S. K. et Shiu, S.C.K., Foundation of Soft case-based reasoning, Wiley Series on Intelligent Systems IEEE, Albus, J.S., Meystel, A.M., Zadeh, L.A., Series Editors, ISBN 0471086355, 9780471086352, 274 pages, 2004.

**[PAV 09]** Pavon F., Diaz R.L. et Luzon V., Automatic parameter tuning with a Bayesian case-based reasoning system. A case of study. Expert Systems With Applications, Vol 36(2P2), pp. 3407-3420, 2009.

**[PEA 88]** Pearl J., Probabilistic reasoning intelligent systems: networks of plausible inference. Morgan kaufman, Palo Alto, 1988.

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

**[QUI 86]** Quinlan J., Induction of decision trees. Machine Learning, Vol 1, pp. 81-106, 1986.

**[QUI 93]** Quinlan J., C4.5, Programs for Machine Learning. Morgan Kaufmann, San Mateo, 1993.

**[RIC 93]** Richter M.M., Wess S., Althoff K.D. et Maurer F., First European Workshop on Case-Based Reasoning, University of Kaiserslautern, Germany, Lecture Notes in Artificial Intelligence, Vol 837, Springer Verlag, Berlin, 1993.

**[RIC 98]** Richter M.M., Introduction In Case-Based Reasoning Technology: From Foundations to Applications. Edited by M. Lena, B. Bartsc-Sporl, H. D. Burkhard, et S.Wess. Springer-Verlag, Berlin, pp. 1-15, 1998.

**[ROS 11]** Rose C., Modélisation stochastique pour le raisonnement médical et ses applications à la télémédecine. Thèse de doctorat, l'université Henri Poincaré, Nancy 1, Mai 2011.

**[SCH 77]** Schank R. et Abelson R., Scripts, Plans, Goals and Understanding. Lawrence Erlbaum Associates, Hillsdale, New Jersey, USA, 1977.

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

**[SCH 05]** Schmidt R. et Gierl L., A prognostic model for temporal courses that combines temporal abstraction and case-based reasoning. International Journal of Medical Informatics, Vol 745(2), pp. 307-315, 2005.

**[SCH 07]** Schmidt R., Case-Based Reasoning in Medicine. Especially an Obituary on Lothar Gierl Studies in Computational Intelligence (SCI) 48, pp. 63-87, Springer-Verlag Berlin Heidelberg, 2007.

**[SCH 09]** Schmidt R. et Vorobieva O., Combining Statistics and Case-Based Reasoning for Medical. Research C.L. Mumford and L.C. Jain (Eds.): Computational Intelligence, ISRL 1, pp. 673-696, Springer-Verlag, Berlin Heidelberg, 2009.

**[SHA 89]** Shafer G. R. et Shenoy P. P., Probability Propagation. In the second International Workshop on Artificial Intelligence and Statistics, Fort Lauder-Dale, Florida, USA, 1989.

**[SIL 00]** Silvia N., Schiaffino S. et Amandi A., User profiling with Case-Based Reasoning and Bayesian Networks. In Open Discussion Track Proceedings, International Joint Conference, IBERAMIA-SBIA, Atibaia, Brazil, pp. 12-21, 2000.

**[SMA 94]** Smail M., Raisonnement à base de cas pour une recherche évolutive d'information. Prototype CARBI-n.vers la définition d'un cadre d'acquisition de connaissance, Thèse de doctorat d'université, Unix. Herni Pointcaré-Nancy I, Octobre 1994.

**[SMY 97]** Smyth P., Belief networks, hidden Markov models, and Markov random fields a unifying view. Pattern Recognition Letters, N° 18, pp. 1261-1268, 1997.

**[SOL 04]** Sollenborn M., Clustering and Case-Based Reasoning for User Stereotypes. Copyright C, ISSN 1651-9256, ISBN 91-88834-66-2, October 2004.

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

**[SOU 02]** Souafi S., Contribution à la reconnaissance des structures des documents écrits: Approche probabiliste. Thèse de doctorat, l'institut national des sciences appliquées de Lyon, 21 Septembre 2002.

**[TRA 08]** Tran H. et Schonwalder J., Fault Resolution in Case-Based Reasoning. In: Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence: Trends in Artificial Intelligence, Springer 429, 2008.

**[TUR 88]** Turner R., Organizing and using schematic knowledge for medical diagnosis. In: Kolodner, J. (Ed.): Proceedings of the Workshop on Case-Based Reasoning, pp. 435-446, Florida, 1988.

**[VEL 95]** Veloso M., Carbonell J., Pérez A., Borrajo D., Fink E. et Blythe J., Integrating Planning and Learning: The PRODIGY Architecture. Journal of Experimental and Theoretical Artificial Intelligence, Vol 7(1), pp. 81-120, 1995.

**[WAT 97]** Watson I., Applying Case-Based Reasoning: Techniques for Enterprise Systems. Morgan Kaufmann Publishers Inc, 1997.

**[WEI 91]** Weiss S.M. et Kulikowski C.A., Computer systems that learn. Artificial intelligence ISSN 0004-3702, Morgan-Kaufmann, 1991, 1993, Vol 62(2), pp. 363-378, 1991.