Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
وزارة التعليم العالي و البحث العلمي

BADJI MOKHTAR-ANNABA UNIVERSITY
UNIVERSITE BADJI MOKHTAR-ANNABA

جامعة باجي مختار - عنابة

Faculté des Sciences de l'Ingénieur **Année : 2014**
Département d'Informatique

THESE

Présentée en vue de l'obtention du diplôme de ***DOCTORAT EN SCIENCES***

Spécialité : ***Informatique***
Option : ***Intelligence Artificielle***

La prise en charge de l'incertain dans le système RespiDiag

Par
GUESSOUM Souad

Directeur de thèse: LASKRI Mohamed Tayeb Pr. Univ. Badji Mokhtar, Annaba

Devant le Jury

Président	M. KIMOUR Mohamed Tahar	Professeur	à l'université d'Annaba
Examinateur	M. KAZAR Okba	Professeur	à l'université de Biskra
Examinateur	M. BABAHENINI Mohamed Chaouki	Maître de conf. A	à l'université de Biskra
Examinateur	M. KHOLLADI Mohamed Khireddine	Professeur	à l'université d'El Oued
Examinateur	M. BOUKERRAM Abdellah	Professeur	à l'université de Béjaia
Invité	M. BENALI Rachid	Professeur	à l'université d'Annaba

Année : 2014

Je dédie ce travail

à la mémoire de mon cher et regrété papa,

à ma chère maman,

à toute ma famille,

et à tous ceux qui m'aiment.

Remerciements

Je tiens à remercier toute personne qui m'a aidée de loin ou de près à réaliser ce travail.

Mes vifs remerciements vont de manière spéciale à mon encadrant Pr LASKRI Mohamed Tayeb, de m'avoir suivie, orientée et encouragée. Qu'il retrouve sur cette page ma vive reconnaissance pour son soutien durant toutes ces années de travail.

Un grand remerciement va également aux membres de jury d'avoir accepter d'examiner ma thèse, de faire partie de mon jury et de m'enrichir ainsi avec leurs remarques et critiques.

Je tiens aussi à remercier les experts médicaux qui m'ont aidée et ont contribué à ce travail grâce à leur savoir-faire, particulièrement le Pr Benali Rachid, qui m'a accueillie dans son service de pneumologie de l'hôpital Dorban d'Annaba. Il a mis à mes dispositions toutes les informations dont le travail avait besoin pour être achevé. Je suis reconnaissante également vis à vis des docteurs Yasmine Moumene et Soumaya Boudraa avec lesquelles j'ai eu de nombreuses et de longues discussions à propos du domaine d'application de notre système.

Je remercie également et énormément Jean Lieber (Maître de conférence à l'université de Henri Poincaré de Nancy 1) pour toutes ses orientations pendant mon stage dans son laboratoire LORIA.

Mon remerciement va aussi au Pr Pierre Spiteri, du laboratoire l'IRIT de Toulouse, pour toute sa gentillesse et sa bonne volonté de m'apporter de l'aide pendant tout mon séjour à Toulouse.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Le raisonnement clinique a été depuis longtemps une source d'inspiration pour l'intelligence artificielle (IA). Il contribue largement dans l'évolution des techniques visant la reproduction du raisonnement humain sur machine. Pour cette reproduction de multiples approches de l'IA ont été exploitées, particulièrement dans la réalisation des systèmes d'aide à la décision. La demande sur ce type de systèmes est plus importante dans les domaines où la connaissance et les expériences évoluent dans le temps de manière rapide, ce qui est le cas du domaine médical.

Depuis de longues années, la médecine a constitué un excellent champ d'expérimentation pour tester les différents paradigmes de l'IA. Ces expérimentations sont très importantes aussi bien pour l'IA que pour la médecine qui prend largement l'intérêt par l'obtention de systèmes d'aide à la décision de diagnostic et de thérapie, des systèmes d'interprétation d'images radiologiques, ainsi que des systèmes de formation. En contre partie, ces systèmes retournent un impact important sur la progression des recherches en IA, parce qu'ils permettent aux chercheurs de tester la validité de leurs approches, de révéler leurs lacunes, d'y apporter correction et donc de mieux évoluer.

Dans les domaines où il est question de prise de décision, l'expérience peut jouer un rôle important dans la résolution de nouveaux problèmes. Ainsi, les solutions des problèmes passés peuvent être utiles pour résoudre un problème courant à chaque fois qu'il y a une ressemblance entre eux. C'est le cas dans le domaine médical où nous observons souvent une ressemblance entre les symptômes des patients présentant la même patholo-

gie, et donc cette similitude de symptômes peut être exploitée pour poser le diagnostic le plus pertinent ou encore proposer la thérapie la plus efficace.

Le raisonnement à partir de cas (RàPC) est une approche de l'IA qui consiste à résoudre un nouveau problème en réutilisant la solution d'un problème passé et similaire. Les problèmes résolus dans le passé forment un ensemble d'expériences qui sont réunies dans la mémoire du système appelée "base de cas". Le cas étant un couple de descriptions des caractéristiques du problème et de sa solution. Le principal avantage de cette technique est qu'elle peut réduire considérablement l'effort d'acquisition de connaissances. En effet, nous n'avons pas besoin de savoir comment l'expert pense pour résoudre le problème, parce que la connaissance pour nous, consiste simplement à établir une description d'un problème et de sa solution généralement sous la forme (attribut, valeur). L'approche se fonde alors sur un grand nombre de problèmes résolus dans le passé au lieu de compter sur les connaissances explicites du domaine.

La litterature peut montrer beaucoup de systèmes RàPC médicaux ayant prouvé leur succès, dû principalement au fait que le raisonnement à parir de cas est très similaire au raisonnement clinique. En effet, durant son activité, le médecin fait souvent appel à sa mémoire pour chercher une certaine ressemblance entre le nouveau problème et les anciens problèmes de ses expériences passées dans l'exercice. Une telle ressemblance (quand elle existe) peut aider énormément dans la résolution du nouveau problème, en terme de faire le diagnostic le plus précis ou encore de proposer le traitement le plus efficace. Le raisonnement à patir de cas permet bien ce recours à l'expérience acquise pour résoudre de nouveaux problèmes, ce qui nous a motivés à l'utiliser dans notre application médicale.

Le premier objectif de cette thèse est de réaliser le système RESPIDIAG [Guessoum11, Guessoum12-a], un système d'aide à la décision pour le diagnostic de la *Broncho Pneumopathie Chronique Obstructive (BPCO)*, une maladie dangereuse, mortelle même cau-

sée par le tabac. Basé sur les principes du raisonnement à partir de cas, notre système recueille l'expérience de plusieurs médecins dans l'objectif d'apporter de l'aide aux futurs pneumologues. Ce travail a été réalisé avec la collaboration des cliniciens spécialistes du service de pneumologie de l'Hôpital Dorban (Annaba, Algérie).

Une phase importante du cycle du raisonnement à partir de cas est la phase de recherche basée sur le calcul des similarités entre les descripteurs du nouveau problème que nous appellerons "problème $cible$" dans la suite de cette thèse et des anciens cas de la mémoire du système que nous appellerons "cas $sources$".

Toutefois, ces descripteurs peuvent contenir des vides dans certains cas. En effet, dans la pratique médicale, le médecin peut être dans une situation au cours de laquelle il doit décider à propos d'un diagnostic alors qu'il n'a pas toutes les données nécessaires. Cela se produit par exemple, dans les services des urgences lorsque le patient arrive en état de crise et le médecin est obligé de diagnostiquer sans avoir le temps d'attendre les résultats des tests médicaux. Ceci fait bien que les dossiers des patients contiennent des données manquantes qui justifient les vides dans la base de cas de RESPIDIAG.

Le deuxième objectif de cette thèse est alors d'intégrer des mécanismes permettant de prendre en charge "l'incertain" dans RESPIDIAG. Nous entendons par cela deux situations différentes : raisonner malgré des données manquantes, ou bien raisonner pour combler les informations manquantes par les valeurs les plus probables.

Lors de la réalisation de RESPIDIAG, nous avons traité ce problème de données manquantes, connu généralement sous le nom du missing data, par la proposition de plusieurs approches que nous avons évaluées par la suite. Ces approches par leurs objectifs différents se distinguent en deux catégories : la première est composée des approches qui raisonnent malgré les données manquantes, elles visent à attribuer une valeur à la similarité entre deux valeurs à chaque fois que l'une de celles-ci est absente.

Ces approches s'exécutent pendant l'état *online* de RESPIDIAG et plus précisément pendant le processus de recherche. Utilisant chacune un principe différent, elles tirent leurs noms de ce principe. Ainsi, nous avons les approches *pessimiste*, avec ses deux variantes *pessimiste*** et *pessimiste**, *médium* avec sa variante *médium**, *sélective* et *optimiste* [Guessoum12-b, Guessoum12-c, Guessoum13].

En outre, le problème des données manquantes a été considéré d'un point de vue hors ligne, et les stratégies proposées dans ce contexte constituent la deuxième catégorie de nos approches [Guessoum14]. Elles visent à combler les vides dans la base de cas avec des valeurs plausibles. Deux méthodes différentes sont utilisées pour estimer ces valeurs. La première consiste à agréger les valeurs connues -de l'attribut en question- dans les autres cas (cette agrégation étant une moyenne des valeurs numériques et le résultat d'un vote pour les valeurs symboliques). Dans la seconde méthode, nous proposons d'utiliser le principe du RàPC lui même pour combler le vide. La solution à trouver étant la valeur manquante.

Ce traitement des données manquantes constitue l'une des contributions de cette thèse. Un autre apport de ce travail vise à faire faire la tâche d'adaptation du cycle de RESPIDIAG par un système expert basé sur un autre mode de raisonnement. Ainsi nous obtenons une fusion du raisonnement à partir de cas avec un raisonnement à base de règles de production [Guessoum06, Guessoum07]. L'objectif étant double,

- exploiter les performances des systèmes experts pour la phase d'adaptation qui est habituellement modélisée par de simples règles dans le système, puis
- fusionner les deux raisonnements qui sont très utilisés par l'homme de manière générale et par le médecin de manière particulière. En réalité le cerveau humain n'est pas limité à un seul raisonnement face à un problème donné. Ainsi, la combinaison des deux modes de raisonnements dans RESPIDIAG le fait rapprocher plus du raisonnement du médecin, et lui permet donc de donner des résultats plus fiables.

La thèse est organisée en trois parties essentielles :

– une première partie, *Etat de l'art*, composée de trois chapitres. Le premier présente le concept du raisonnement à partir de cas, le deuxième présente les systèmes experts à base de règles de production. Et le troisième est consacré aux travaux de la littérature en relation avec les apports de RESPIDIAG.
– une deuxième partie, *Le système* RESPIDIAG , consacrée aux détails conceptuels de notre système et composée aussi de trois chapitres. Le premier introduit le domaine d'application, l'architecture de RESPIDIAG en plus des définitions du problème et de sa solution, ainsi que les différentes notations utilisées dans la thèse. Le deuxième chapitre quant à lui est consacré aux différentes approches proposées pour la phase de recherche, alors que le troisième détaille notre apport pour la phase d'adaptation.
– une troisième partie, *Implémentation et évaluations*, composée à son tour aussi de trois chapitres. Le premier introduit les données de la base de cas, les données de l'échantillon de test et les différentes méthodes d'évaluations utilisées. Les deux autres détaillent les résultats de nos implémentations au niveau des deux phases de recherche et d'adaptation. Ensuite une discussion est donnée, et une conclusion avec des perspectives viendront clôturer cette thèse.

Première partie

Etat de l’art

Chapitre 1

Le raisonnement à partir de cas

Sommaire

1.1 Introduction

L'un des buts principaux de l'intelligence artificielle est de concevoir des systèmes informatiques capables de reproduire les différents modes du raisonnement de l'homme, parmi lesquels nous pouvons distinguer le raisonnement déductif qui déduit de nouvelles connaissances à partir de celles déjà acquises, le raisonnement inductif qui généralise une idée à partir d'observations effectuées, le raisonnement adductif qui recherche les causes d'un fait, et le raisonnement *par analogie* qui interprète une nouvelle situation par comparaison avec une situation voisine.

1.2 Le raisonnement par analogie

Le raisonnement par analogie est reconnu être très utilisé par l'homme. En effet, et face à une situation donnée, l'expérience d'une situation semblable peut s'avérer très utile. Ce raisonnement est appelé principalement dans deux contextes différents :

- Quand la résolution du problème s'annonce particulièrement complexe et longue. L'utilisation de la solution d'un problème similaire déjà résolu permet alors 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.

Observons la figure 1.1 connue sous le nom du carré d'analogie. 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 ?.

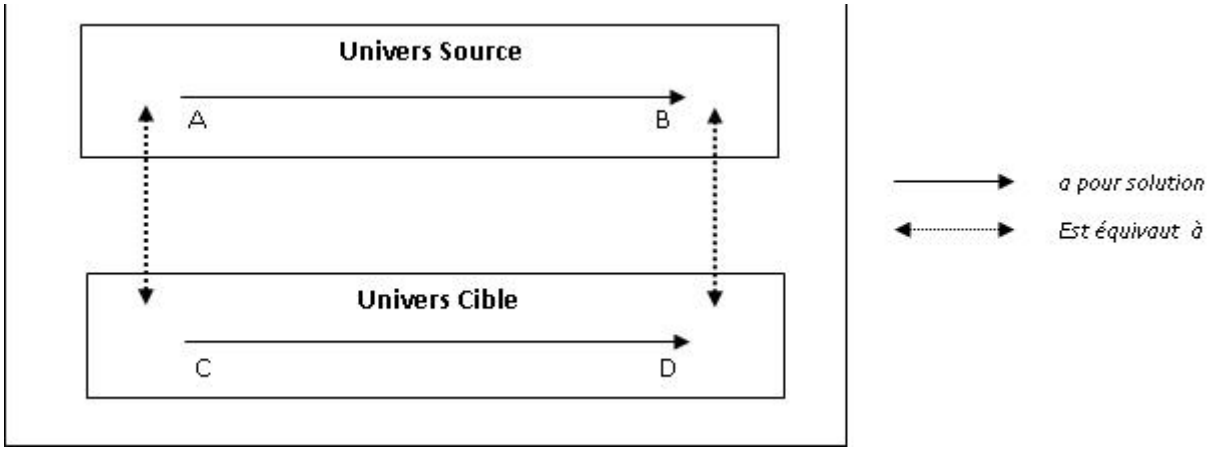

FIGURE 1.1 – le carré d'analogie.

1.3 Le principe du RàPC

Le RàPC est une technique basée principalement sur le raisonnement analogique. Elle a été inspirée d'un réel modèle congnitif observé dans le comportement humain, qui consiste à apprendre à partir des expériences passées pour apporter une solution à un nouveau problème. En effet, il a été démontré en psychologie que dans beaucoup de situations, l'homme commence à résoudre ses problèmes en se basant plus sur ses propres expériences que sur les connaissances du domaine. Donc l'idée globale du RàPC consiste à réutiliser la solution d'un ancien problème similaire pour résoudre un nouveau. Toutes les expériences passées sont réunies dans une mémoire appelée "base de cas", où le cas est un couple de descripteurs du problème et de sa solution. A ces deux élements basiques, vient des fois s'ajouter un troisième descripteur apportant une explication, une justification ou encore une estimation de la réutilisation du cas.

Nous donnons les définitions suivantes :

- *Un cas source* : est une description d'un problème préalablement analysé, résolu et structuré de façon à pouvoir l'adapter dans la résolution de futurs problèmes. Il est mémorisé dans la base de cas et il est défini comme étant le couple d'énoncés problème-solution (p, sol(p)).
- *Un problème cible* : c'est la description d'un nouveau problème à résoudre, auquel on doit trouver une solution en se basant sur les cas sources. On l'appelle également cas cible.
- *Une base de cas* : est un ensemble fini de cas résolus, base-de-cas = {(pi, sol(pi) / i=1..n}. Elle constitue la mémoire d'un système RàPC. L'ensemble des cas sources sont stockés dans cette base de cas dont l'organisation est très importante pour le bon déroulement du processus de recherche. Il existe plusieurs types d'organisation de la mémoire, la plus célèbre étant l'organisation plate.

1.4 Les connaissances dans un système RàPC

Les connaissances utilisées pour concevoir un système RàPC se divisent en quatre catégories :

- *Le vocabulaire d'indexation* : est un ensemble d'attributs qui caractérise la description des problèmes et des solutions du domaine.
- *La base de cas* : c'est l'ensemble des expériences structurées qui seront exploitées par les phases de recherche, d'adaptation et de maintenance.
- *Les mesures de similarité* : qui sont un ensemble de fonctions utilisées pour évaluer la similarité entre les cas. Ces mesures sont définies selon des faits et sont exploitées pour la phase de recherche dans la base de cas.
- *Les connaissances d'adaptation* : qui sont des heuristiques généralement sous forme de règles, permettant de modifier les anciennes solutions et d'évaluer leur applicabilité à de nouvelles situations.

1.5 Le cycle du raisonnement à partir de cas

Le cycle du RàPC est identifié par Aamodt et Plaza dans [Aamodt94] comme étant un processus de quatre étapes.

- *La phase de recherche (ou de remémoration)* : C'est la phase la plus importante du cycle. Elle consiste à mesurer la similarité entre le problème courant et les anciens cas rassemblés dans la mémoire du système pour récupérer le (ou les) cas le(s) plus similaire(s). Ce processus est basé sur les métriques de similarité dont le choix est fondamental. En effet, pour un cas cible donné, la bonne sélection du cas le plus similaire dépend strictement de ces métriques. Il existe de multiples mesures qui peuvent être utilisées dans cette phase, la plus populaire étant la méthode du k-plus proches voisins.

– *La phase d'adaptation (ou de réutilisation)* : c'est l'étape la plus délicate, elle permet d'adapter la solution (s'il y a besoin) du cas récupéré de la phase précédente aux données du problème cible. Sa difficulté est principalement dûe au fait que les heuristiques et les connaissances utilisées dans cette phase dépendent strictement du domaine d'application. Pour certains systèmes RàPC particulièrement ceux dédiés au diagnostic, cette phase est généralement ignorée, et le processus consiste juste à trouver le diagnostic le plus proche.
– *La phase de révision* : durant laquelle, la solution adaptée est présentée à l'utilisateur qui décidera à propos de sa validité. Dans le cas affirmatif la dernière phase est entamée, sinon il peut procéder à une modification de la solution.
– *La phase d'apprentissage* : elle consiste à ajouter le nouveau problème avec sa solution dans la base de cas, ce qui donne un nouveau cas appris.

La figure 1.2 schématise les différentes phases du cycle du RàPC.

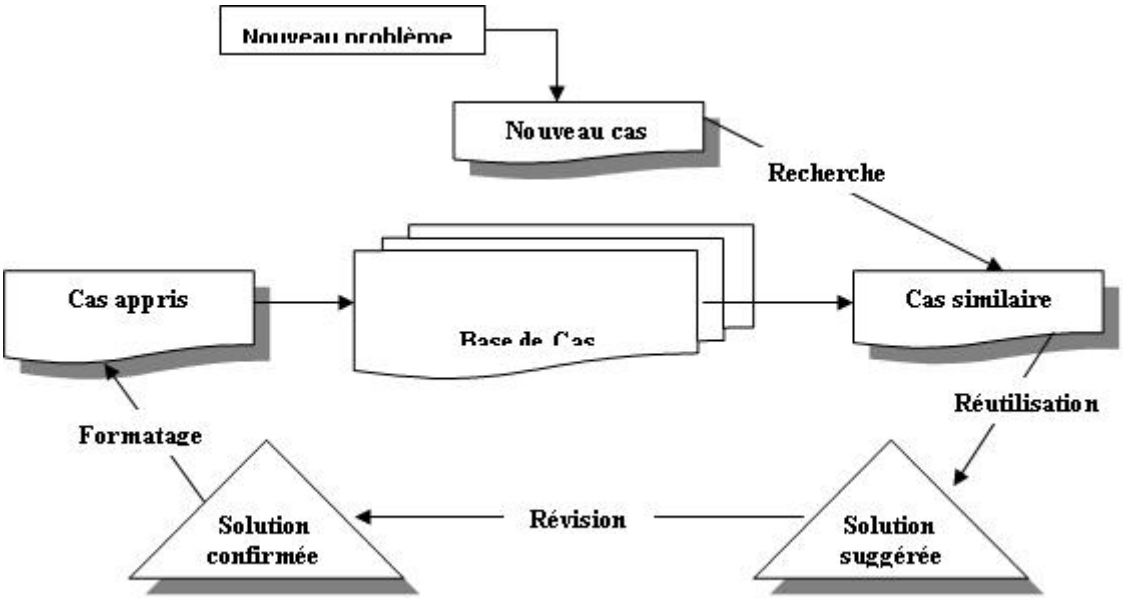

FIGURE 1.2 – Le cycle du raisonnement à partir de cas

1.6 Avantages et inconvénients du RàPC

Le RàPC présente plusieurs avantages dont le plus important est le fait qu'il ne nécessite pas une acquisition de connaissances "approfondies" du domaine considéré. En effet, nous n'avons pas besoin de savoir comment l'expert pense pour résoudre son problème, parce que la modélisation de la connaissance consiste juste à établir une description du problème et de sa solution. Donc le problème d'acquisition de connaissances issues d'un consensus entre experts du domaine est évité. Le second avantage est que l'implémentation de cette technique est relativement simple comparée à l'implémentation des autres techniques de l'intelligence artificielle. Le troisième point de force de ce mode de raisonnement est que l'apprentissage est facilement réalisé dans les systèmes RàPC. Il consiste souvent et simplement en l'insertion des nouveaux cas dans la mémoire du système. Un autre point fort des systèmes RàPC est leur capacité d'apprendre de leurs échecs. Cela est dû au faite qu'ils permettent aussi de mémoriser les causes d'échec si une tentative de résolution échoue.

Parallèlement à ces avantages, le RàPC souffre de quelques difficultés, notamment au niveau des deux phases de recherche et d'adaptation.

D'après Mantaras et Plaza dans [Mantaras96], le problème de base dans les systèmes RàPC est bien la recherche, et donc la sélection des cas semblables. Certainement parce que le succès des phases suivantes, à savoir l'adaptation et la révision, dépend étroitement de la pertinence des cas sélectionnés lors de cette première phase. Une bonne indexation s'impose et le choix des attributs sur lesquels se fait la recherche doit être minutieux. Aussi, le choix d'une bonne métrique de similarité est déterminant pour la performance du système. C'est la raison pour laquelle, une part importante de travaux de recherche s'est focalisée sur les mesures de similarité les plus génériques possibles [Mille99]. Nombreuses heuristiques ont été développées pour rendre cette recherche la plus fiable possible surtout que le bon déroulement des phases suivantes en est très dépendant.

Sur un autre plan, il est fréquent que les attributs décrivant le problème ne participent pas tous avec la même importance dans la détermination de la solution. Cet aspect est pris en charge par l'affectation de coefficients de pondération à chaque attribut. En pratique, les valeurs de ces poids ne sont pas évidentes à fixer au début du travail et peuvent être ajustées pendant les épisodes d'essai pour s'approcher des valeurs qui donnent les meilleurs résultats. La recherche est basée donc sur le calcul des similarités pondérées.

Tandis qu'au début des années 90, les chercheurs se focalisaient sur les tâches de recherche et du calcul de similarité, vers la fin de cette période, ils se sont investis dans l'étude des divers aspects de l'adaptation qui est un processus très compliqué. Donc une autre difficulté dont souffre le RàPC se situe au niveau de cette phase. Bien que des efforts ont été développés pour résoudre le problème de l'adaptation, ce dernier demeure encore résistant. Certainement parce qu'il est très dépendant du domaine d'application et il ne peut être résolu de manière généralisée. Dans l'objectif de surmonter ce problème de dépendance, et dans un projet assez ambitieux Fuchs et Mille ont proposé dans [Fuchs00] un modèle générique pour la tâche d'adaptation en 1998 et l'ont complété en 2000 par des modélisations au niveau connaissance pour le cas et pour la tâche de remémoration.

Chapitre 2

Les systèmes experts

Sommaire

2.1 Introduction

Un système expert que nous noterons plus simplement dans ce qui suit par "SE", est un programme informatique capable de résoudre des problèmes dont la solution requiert une expertise humaine considérable dans un domaine qui ne se prête pas à la programmation algorithmique. Il possède une grande masse de connaissances qu'il acquit d'un expert humain du domaine.

2.2 Le raisonnement à base de règles de production

Le raisonnement à base de règles de production est le raisonnement le plus ancien et le plus répondu dans le monde de l'IA. Un système expert à base de règles de production emploie la connaissance du domaine formalisée sous forme de règles de production composées de deux parties : une *condition* et une *conclusion*. Chaque règle de production représente une unité de la connaissance liée au champ d'étude, et se présente sous la forme :

Si *condition* **alors** *conclusion*

La partie *condition* est composée d'une ou de plusieurs prémisses reliées entre elles par des conjonctions (et) ou des disjonctions (ou) ou les deux en même temps. La partie *conclusion* regroupe toutes les conséquences, conclusions ou actions qui doivent être réalisées lorsque la partie *condition* est vérifiée (évaluée à "vrai" compte tenu des éléments de la base de faits).

2.3 Les composantes d'un système expert à base de règles

Un système expert à base de règles de production est composé de deux parties indépendantes :

- *une base de connaissances (BC)* elle-même composée : d'une *base de règles (BR)* qui modélise la connaissance du domaine considéré, et d'une *base de faits (BF)* qui contient les informations concernant le problème à résoudre. Ces deux bases sont très liées au domaine d'application.
- *un moteur d'inférences (MI)* capable de raisonner et de faire des déductions à partir des informations contenues dans la *BC*. Il est totalement indépendant du domaine d'application.

A ces deux composantes de base, viennent s'ajouter généralement deux autres éléments :

- *une interface expert* qui permet de mettre à jour la *BR* au fil des temps et,
- *une interface utilisateur* qui permet les intéractions entre le système et l'utilisateur. Elle doit être la plus conviviale possible, de façon à rendre parfaitement aisée l'introduction de la part de l'utilisateur des différents faits relatifs au problème à résoudre, et l'explication des raisonnements par le système.

L'architecture d'un système expert peut être donc illustrée par la figure 2.1.

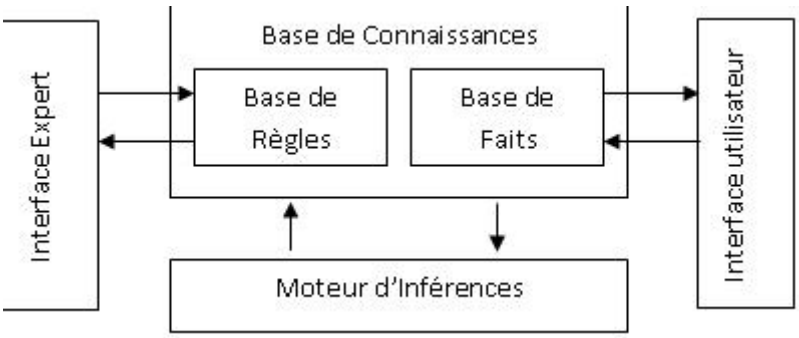

FIGURE 2.1 – Architecture d'un SE.

L'indépendance entre la *BC* et le *MI* est un élément essentiel des systèmes experts. Elle permet une représentation des connaissances sous forme purement déclarative. Autrement dit, sans aucun lien avec la manière dont ces connaissances sont utilisées. L'avantage de ce type d'architecture est qu'il est possible de faire évoluer les connaissances du système sans avoir à agir sur le mécanisme de raisonnement. Il en est de même pour nous les humains : un accroissement ou une modification de nos connaissances n'entraîne pas nécessairement une restructuration en profondeur de nos mécanismes de raisonnement.

2.4 Les modes de fonctionnement d'un MI

Un moteur d'inférence est un mécanisme qui permet d'inférer des connaissances nouvelles à partir de la base de connaissances du système. On distingue essentiellement trois modes principaux de fonctionnement des moteurs d'inférences : le chaînage avant, le chaînage arrière et le chaînage mixte.

– *Le chaînage avant*, son mécanisme est très simple : pour déduire un fait particulier, on déclenche les règles dont les prémisses sont connues jusqu'à ce que le fait à déduire soit également connu ou qu'aucune règle ne soit plus déclenchable.

– *Le chaînage arrière*, son mécanisme consiste à partir du fait que l'on souhaite établir, à rechercher toutes les règles qui concluent sur ce fait, à établir la liste des faits qu'il suffit de prouver pour qu'elles puissent se déclencher puis à appliquer récursivement le même mécanisme aux faits contenus dans ces listes. L'algorithme de chaînage arrière est nettement plus compliqué que le précédent, et il présente en plus l'inconvénient qu'il peut boucler si les faits déjà examinés ne peuvent pas être mémorisés (par exemple parce qu'ils sont trop nombreux).

– *Le chaînage mixte*, dont l'algorithme combine comme son nom l'indique, les algorithmes de chaînage avant et de chaînage arrière. Son principe est de déduire tous les faits qui peuvent être déduits par le chaînage avant, puis appliquer le principe de chaînage arrière pour déduire d'éventuels nouveaux faits, à partir du fait que l'on souhaite obtenir, et les ajouter à la base de faits. Dans certaines situations, l'utilisation de ce type de chaînage s'avère nettement meilleure.

Chapitre 3

Travaux en relation avec RESPIDIAG

Sommaire

3.1 Introduction

Dans ce chapitre, nous donnons un aperçu sur certains systèmes que nous avons jujés proches de RESPIDIAG. Ce dernier comme nous l'avons précisé au niveau de l'introduction de cette thèse, est un système RàPC dédié au diagnostic médical, qui intègre un système expert dans son processus, et qui traite de certaines manières proposées le problème du missing data. Les systèmes que nous avons sélectionnés dans ce chapitre peuvent être en relation avec RESPIDIAG,

- parce qu'ils sont basés sur le principe du RàPC et sont dédiés aussi au diagnostic médical, ou bien,
- parce qu'ils sont des systèmes experts basés sur les règles de production et dédiés

au diagnostic médical, ou bien,

– parce que ces systèmes traitent d'une manière ou d'une autre le problème des données manquantes.

3.2 Des systèmes RàPC médicaux

La litterature montre beaucoup de systèmes RàPC médicaux ayant prouvé leurs succès. Rainer Schmidt présente dans [Schmidt07] un aperçu sur les systèmes RàPC médicaux réalisés dans les vingts dernières années. Il distingue principalement trois tendances.

Nous trouvons en premier lieu les systèmes d'aide aux diagnostics tel que : TeComMed [Schmidt01] qui fournit des prévisions sur les épidémies de grippe, CASEY [Koton88] dédié au diagnostic des troubles cardiaques, FLORENCE [Bradburn93] dédié aux diagnostic, pronostic et prescription de soins infirmiers. Ici, le diagnostic n'est pas utilisé dans le sens médical commun de l'identification d'une maladie, mais il cherche à répondre à la question «quel est l'état de santé actuel d'un patient en requête ?. Nous trouvons également MERSY [Opiyo10] qui prend en charge la santé des travailleurs dans les régions rurales et FM-Ultranet [Balaa03] pour le diagnostic des déformations du foetus.

La seconde tendance est tracée selon toujours [Schmidt07] par les systèmes thérapeutiques, tel que : ICONS [Schmidt07] qui fournit des conseils dans le traitement antibiotique, ISOR [Schmidt09] qui est un système intéractif aidant les médecins dans l'explication et l'interprétation des positions exceptionnelles, où l'approche thérapeutique théorique ne donne pas le résultat estimé dans des cas particuliers. Nous trouvons également TA3-IVE [Jurisica98] qui fournit des planings pour le traitement des cas de fécondation in-vitro, et KASIMIR [Lieber00] qui propose des traitements pour le cancer du sein.

Dans la troisième tendance, nous trouvons des systèmes spécialisés dans l'interprétation des images médicales, tel que : MacRad [Gierl98] et SCINA [Haddad97].

Bichindaritz présente aussi dans [Bichindaritz10], une synthèse sur les applications de l'approche RàPC dans les sciences de la santé. Beaucoup d'exemples de systèmes médicaux dédiés à la tâche de diagnostic sont donnés, comme les systèmes SHRINK (psychiatrie, 1987), PROTOS (troubles auditoires, 1987), CASEY (troubles cardiaques, 1988), et MNAONIA (psychiatrie, 1994). A une longue liste, viendra alors s'ajouter notre système RESPIDIAG (BPCO, 2013).

L'ensemble de ces systèmes est synthétisé dans la table 3.1.

N	Nom du système	Objectif	Année
1	SHRINK	diagnostic des troubles psychiatriques	1987
2	PROTOS	diagnostic des troubles auditoires	1987
3	CASEY	diagnostic des troubles cardiaques	1988
4	Florence	diagnostic, pronostic et prescription de soins infirmiers	1993
5	MNAONIA	diagnostic des troubles psychiatriques	1994
6	MERSY	prise en charge de la santé des travailleurs dans les régions rurales	1995
7	SCINA	interprétation d'images médicales	1997
8	TA3-IVE	planing de traitement pour des cas de fécondatio in-vitro	1998
9	MacRad	interprétation d'images médicales	1998
10	KASIMIR	traitement des cancers du sein	2000
11	TeComMed	prévisions sur les épidémies de grippe	2001
12	FM-Ultranet	diagnostic des déformations du foetus	2003
13	ICONS	traitement antibiotique	2007
14	ISOR	interprétations des positions exceptionnelles	2009

TABLE 3.1 – Synthèse des systèmes RàPC de notre état de l'art

3.3 Des systèmes experts médicaux

Le raisonnement basé sur les règles de production a été dans les années 70 le choix le plus populaire des chercheurs pour établir des systèmes experts au profit de la médecine. Ces chercheurs étaient plus attirés par la réalisation d'outils d'aide au diagnostic et à la thérapie que par d'autres aspects de l'activité médicale. Ceci explique le fait que presque la totalité des systèmes experts médicaux réalisés dans cette époque était dédiée au diagnostic [Chae96]. Et au cours des trente dernières années, beaucoup d'autres systèmes experts médicaux ont été développés. Les motivations ayant été nombreuses ! Il s'agit principalement :

- d'améliorer l'exactitude du diagnostic par des approches systématiques,
- d'améliorer la fiabilité des décisions en évitant des influences sans garantie des cas semblables mais non identiques,
- d'approfondir la compréhension de la structure de la connaissance médicale et,
- d'améliorer la compréhension du processus de prise de décision clinique afin d'améliorer l'enseignement médical.

Les systèmes médicaux à base de règles de production sont très nombreux dans la littérature. Nous nous limitons ci-dessous à la citation des plus connus.

MYCIN, est le plus célèbre. Il est dédié au diagnostic des infections du sang.

INTERNIST quant à lui, est dédié au diagnostic des maladies liées au champ de la médecine interne. Apparut en 1982, il fonctionne sur une base de connaissances plus volumineuse que celle de MYCIN car il traite un grand nombre de maladies (environ des centaines). Il est connu actuellement sous le nom de CADUCEUS. Sa force est qu'il considère toutes les combinaisons possibles de pathologies chez un patient. Cette force même a conduit par la suite à un inconvénient majeur qu'est l'explosion combinatoire de la base de connaissance.

QMR (Quick Medical Reference) apparut en 1988, est une version simplifiée et plus didactique de INTERNIST. Il utilise la base de connaissances d'INTERNIST, et aide dans le diagnostic en fonction des symptômes du patient, des résultats d'examen, et des essais en laboratoire. Il incorpore plus de 4000 manifestations possibles des maladies et offre une fiabilité dans les diagnostics à un niveau comparable aux médecins praticiens.

ICÖNE qui aide des radiologistes avec le processus du diagnostic différentiel de l'affection pulmonaire vue sur des radiographies de poitrine [Kahn94].

DIATELIC [Thomesse04] : développé par Thomesse et son équipe. C'est un système intelligent de télémédecine appliqué à la dialyse et initié en 1999. Il comporte plusieurs sous systèmes intelligents. Le premier était basé sur des règles de production, succédé par une modélisation de l'état d'hydratation avec le modèle markovien. Un troisième sous-système a été intégré par la suite basé sur un modèle bayésien. Son objectif était d'expliciter l'influence d'une variable aléatoire sur une autre, par exemple l'influence du taux d'hydratation sur le poids réel.

3.4 Des systèmes ayant géré le missing data

Le problème du missing data a fait coulé beaucoup d'encre. Des travaux multiples ont été réalisés pour y remédier utilisant chacun une méthode ou un principe différent, tels que les réseaux bayésiens, les réseaux de neuronnes.... Nous allons évoquer quelques uns dédiés au diagnostic médical.

Dans [Lin08], les auteurs utilisent la technique du réseau bayésien pour proposer un système d'aide à la décision médicale appliqué au diagnostic de 5 maladies : la pancréatite aiguë, l'insuffisance rénale aiguë, l'asthme, la pneumopathie, et les infections des voies urinaires. Dans ce travail, et pour gérer le problème du manque de données, les auteurs proposent quatre méthodes pour préparer les données cliniques à faire entrer dans leur réseau bayésien. Ils supposent que l'absence de données peut être en elle même porteuse

d'information. Et ils décrivent une nouvelle approche qui est d'utiliser l'information représentée par les données absentes pour apporter de l'aide au diagnostic de pathologies.

Leur première méthode consiste simplement à faire l'apprentissage du réseau bayésien sans pré-traitement des valeurs manquantes, la seconde méthode consiste à calculer la valeur manquante par la moyenne globale de toutes les valeurs disponibles dans l'ensemble de données. Leur troisième approche consiste à donner une valeur symbolique au lieu de valeur numérique. Une valeur `missing` lorsque la valeur est inconnue et une valeur correspondant à un intervale numérique à la place. La quatrième stratégie consiste à ajouter de nouvelles variables logiques à l'ensemble de données. Ces variables sont utilisées pour indiquer la présence/absence des valeurs de variables correspondantes qui pourraient être manquantes. Les auteurs précisent que ce traitement de données manquantes a bien amélioré la qualité des résultats de leur application.

Dans [Pesonen98], un autre travail traitant le problème des données manquantes peut être trouvé. Les auteurs proposent un système d'aide à la décision médicale dédié au diagnostic de la douleur abdominale aigue basé sur les techniques des réseaux de neuronnes. Le problème de l'absence de l'information considéré dans ce système est à propos d'un attribut spécifique : le nombre des leucocytes. Les quatres approches utilisées sont : les moyens de substitutions, la méthode aléatoire basée sur la distribution des valeurs connues, la méthode du plus proche voisin, et la méthode des réseaux de neuronnes. Les différentes méthodes montrent de grands écarts dans les valeurs substituées, à l'exception des deux dernières, qui convergent dans la majorité des cas. Le réseau de neurones a amélioré le résultat de leur système à une précision de diagnostic atteignant les 78% (à comparer avec 74% quand la méthode aléatoire basée sur la distribution des valeurs connues a été utilisée).

Dans [Nuovo11], un autre travail traitant le missing data peut être trouvé. Il est dédié au diagnostic du retard mental chez l'adulte et l'enfant. Ce travail considère une base

de données (contenant des vides) de patients diagnostiqués comme affectés par un retard mental léger ou modéré. L'auteur utilise la méthode de l'analyse discriminante pour classifier les patients et comparer les résultats de plusieurs méthodes pour la complétion des données. Il s'agit de :

- l'estimation par la régression, qui utilise la régression linéaire pour obtenir une estimation des valeurs manquantes,
- l'estimation par la maximisation d'espérence, qui est basée sur l'algorithme EM, qui consiste en deux phases : la phase "E"' (expectation) prédicte des valeurs initiales pour les valeurs manquantes utilisant d'autres méthodes (e.g. régression linéaire multiple), et la phase "M" (maximization) où les valeurs manquantes sont calculées itérativement utilisant la fonction de vraisemblance maximale jusqu'à ce que la précision désirée est atteinte.
- la méthode de l'algorithme du clustering flou : (fuzzy C-Means) qui a amélioré le résultats (avec 82.8% de réponses correctes de classifications de cas), relativement aux deux méthodes statistiques.

Deuxième partie

Le système RESPIDIAG

Chapitre 4

Introduction

Sommaire

4.1 Le domaine d'application : la BPCO

Dans cette thèse nous présentons les détails de notre système d'aide à la décision : RESPIDIAG. Il est dédié au diagnostic de la broncho-pneumopathie chronique obstructive qui est une obstruction des branches pulmonaires dûe à une inflammation causée par le tabac [Fabbri11]. La majorité des personnes affectées par cette maladie sont des fumeurs ou ex-fumeurs agés de plus de 50 ans. Cette maladie est très dangereuse, mortelle même : l'Organisation Mondiale de la Santé a estimé à 2.74 millions le nombre de décès dûs à la BPCO en 2000. Aujourd'hui elle est la 4^{eme} cause mondiale de décès alors qu'elle occupait la 6^{eme} place en 1990 [Sweb].

La BPCO dans son état de base peut évoluer à travers quatre stades de sévérité [Anane04]. Elle est au stade I quand les symptômes sont épisodiques, au stade II lorsque

les symptômes sont permanents, au stade III quand la dyspnée apparait même au repos, et au stade IV lorsqu'on observe une incapacité respiratoire. Dans RESPIDIAG nous considérons la BPCO dans son état de base avec ses quatre stades de sévérité, en plus de deux autres exacerbations possibles de la BPCO, qui peuvent être d'origine infectieuse ou d'origine pneumo-thorax. Donc le diagnostic vise à catégoriser le patient dans l'une de ces six classes de diagnostics.

Un diagnostic de BPCO (à son état de base ou dans ses deux états d'exacerbation) est basé sur une longue liste de symptômes, à savoir :

- l'age,
- la température,
- la toux, qui peut être sèche ou productive, avec une expectoration mucqueuse ou purulente,
- la dyspnée,
- le volume expiratoire maximal à la première seconde (VEMS),
- la spirométrie dont le test peut révéler un trouble ventilatoire obstructif (TVO) réversible ou irréversible, un trouble ventilatoire restrictif (TVR), ou bien absence de trouble (état normal),
- le débit expiratoire de pointe (DEP),
- les problèmes antécédents, où nous cherchons si le patient a déjà vécu des crises de dyspnées dans son passé, et si elles ont été suivies ou non,
- la distension thoracique,
- la clarté, qui est une donnée radiologique,
- l'opacité, qui est une donnée radiologique aussi
- la douleur thoracique,
- les râles sibilants,
- l'hyperleucocytose, une donnée que nous obtenons après analyse de sang, elle affirme la présence d'une infection,

- le tabagisme, qui peut être actif, passif ou absent,
- le nombre de paquets de cigarettes fumées par année (en moyenne) et,
- la profession, qui est utilisée car les personnes travaillant dans des milieux pollués sont plus exposées à une atteinte par la BPCO qu'on appelle "professionnelle" dans ce cas.

4.2 Architecture de RESPIDIAG

L'architecture de RESPIDIAG obéit aux besoins de son processus de raisonnement à partir de cas. En effet, elle est composée :

- d'une interface utilisateur qui permettra d'introduire au système les différents symptômes du patient,
- d'un module qui se charge de la tâche de remémoration, son rôle est de sélectionner depuis la base de cas du système, le cas le plus proche du nouveau patient,
- d'un autre module spécifique pour la tâche d'adaptation,
- d'un troisième module pour la phase de révision,
- d'un quatrième qui s'occupera de la tâche d'apprentissage et
- En plus évidemment de la base de cas constituant la mémoire du système qui rassemble toutes les expériences passées (les dossiers archives des patients) sous forme de cas.

La contribution de cette thèse est au coeur des deux premiers modules du cycle RàPC, à savoir la recherche et l'adaptation. Les deux derniers étant faits de manière assez simple dans RESPIDIAG, peuvent être détaillés ou développés dans des travaux ultérieurs.

La deuxième phase du cycle de raisonnement à partir de cas de RESPIDIAG est modélisée par un module spécifique : un système expert à base de règles de production. Son rôle serait d'assurer la tâche d'adaptation de la solution retenue de la phase précédente aux données (symptômes) du nouveau problème (patient). L'idée de fusionner les deux types

de raisonnement a été motivée par le fait que ce sont les plus utilisés par l'homme dans sa vie courante. Ils sont souvent présents tous les deux dans la même prise de décision par l'homme de manière générale et par le médecin de manière particulière.

Notons également que les systèmes experts basés règles ont prouvé leur performance dans la modélisation du raisonnement qui peut être expliqué et formalisé par des règles. Ceci encourage bien de l'utiliser et de le tester à travers une tâche très délicate du cycle du RàPC qui est habituellement réalisée par de simples règles dans le système.

L'architecture du système RESPIDIAG et son cycle sont donnés dans la figure 4.1.

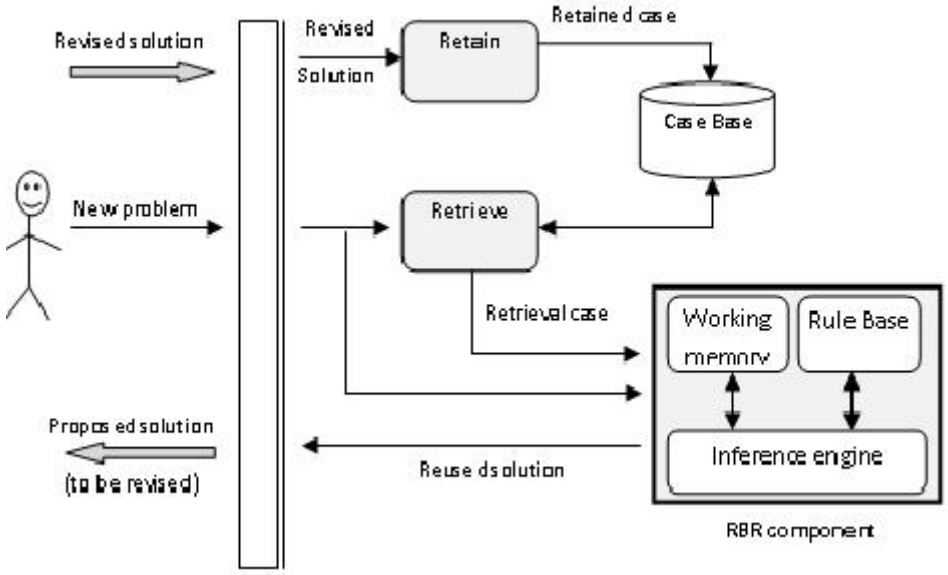

FIGURE 4.1 – Architecture de RESPIDIAG

4.3 La représentation des cas dans RESPIDIAG

RESPIDIAG possède une base de cas structurée en mémoire linéaire, notée par `BaseDeCas` dans ce qui suit. C'est l'ensemble de nos cas sources stockés suivant une représentation <attribut-valeur> qui nous a parut la plus appropriée. Cette dernière est basée sur un ensemble de 17 attributs sur lesquels est fondé le diagnostic et qui ne sont que

les symptômes descripteurs de l'état du patient à son arrivée au service et/ou pendant son séjour hospitalier.

Un cas de `BaseDeCas` est défini par un couple ordonné $(\texttt{pb}, \texttt{sol}(\texttt{pb}))$ où `pb` est un problème décrivant les conditions du patient et $\texttt{sol}(\texttt{pb})$ est le diagnostic solution associé à `pb` :

- `pb` est décrit alors par les 17 attributs correspondant aux symptômes présentés dans la section 4.1, et qui auront dans la suite de ce travail les notations suivantes : `age`, `température`, `toux`, `dyspnée`, `VEMS`, `spirometrie`, `DEP`, `antécédents`, `distensionThoracique`, `hyperLeucocytose`, `clarté`, `opacité`, `douleurThoracique`, `râlesSibilants`, `tabagisme`, `nombrePaquetParAnnée` et `profession`.

  La majorité de ces attributs portent un sens implicite, les autres ont été expliqués dans la section 4.1.

  L'ensemble de ces 17 attributs est noté dans ce qui suit par `Attributs`. La plage des données d'un attribut a est notée par $\texttt{plage}(a)$.

- $\texttt{sol}(\texttt{pb})$ est l'un des six diagnostics du champ d'étude et qui seront notés dans ce qui suit par : `Stade_`k ($k \in \{\texttt{I}, \texttt{II}, \texttt{III}, \texttt{IV}\}$) pour BPCO stade k, `ExInfec` pour exacerbation de BPCO d'origine infectieuse, et `ExPnThor` pour exacerbation de BPCO d'origine pneumo-thorax.

Les valeurs des attributs du problème sont de différents types :

- numérique pour :`age`, `température`, `nombrePaquetParAnnée`, `VEMS`, et `dyspnée` [1],
- logique pour : `DEP`, `distensionThoracique`, `râlesSibilants`, `clarté`, `hyperLeucocytose`, `opacité`, `douleurThoracique`, et

1. La plage de données de l'attribut dyspnée donne les différents stades de sa sévérité, elle est $\texttt{plage}(\texttt{dyspnée}) = \{0, 1, 2, 3, 4, 5\}$, où 0 représente l'absence de dyspnée, et 5 représente le stade le plus sévère de la dyspnée

– symbolique (i.e., types énumérés) pour : `toux` (5), `antécédents` (5), `spirometrie` (4), `tabagisme` (3) et `profession` (20) ;

Les nombres entre parenthèses sont les nombres de valeurs possibles de l'attribut. par exemple, les plages de données de `toux` et `spirometrie` sont :

$$\begin{aligned}\texttt{plage}(\texttt{toux}) = \{&\texttt{pasDeToux}, \texttt{sèche},\\ &\texttt{PM} \text{ (productive avec expectoration muqueuse)},\\ &\texttt{PP} \text{ (productive avec expectoration purulente)},\\ &\texttt{PMP} \text{ (productive avec expectoration muco-purulente)}\}\end{aligned}$$

$$\begin{aligned}\texttt{plage}(\texttt{spirometrie}) = \{&\texttt{normal},\\ &\texttt{TVOR} \text{ (Trouble ventilatoire obstructif réversible)},\\ &\texttt{TVOI} \text{ (Trouble ventilatoire obstructif irréversible)},\\ &\texttt{TVR} \text{ (Trouble ventilatoire restrictif)}\}\end{aligned}$$

Etant donné un problème `pb` et un attribut a, $\texttt{pb}\,.\,a$ est la valeur de l'attribut a de `pb`. Par exemple, si `pb` est tel que

$$\texttt{pb.age} = 60 \qquad \texttt{pb.râlesSibilants} = \texttt{faux} \qquad \texttt{pb.toux} = \texttt{sèche}$$

alors `pb` présente un patient de 60 ans, sans râles sibilants et avec une toux sèche.

Finalement, un cas source est un cas de la base de cas, et est noté par $(\texttt{srce}, \texttt{sol}(\texttt{srce}))$, et un problème cible est noté par `tgt` (pour target).

La procédure de recherche de RESPIDIAG, comme est le cas dans tous les systèmes RàPC, est basée sur une mesure de similarités globale $\mathcal{S}$ entre `srce` et `tgt`, basée à son tour sur d'autres mesures de similarités locales $\mathcal{S}_a$ associées à chaque attribut a.

Dans le chapitre suivant, nous présentons tous les détails se rapportant au processus de recherche de RESPIDIAG, avec la principale contribution de cette thèse consistant à gérer les données manquantes dans la base de cas et dans le problème cible.

Ce chapitre sera suivi par un autre traitant la phase la plus délicate du processus de raisonnement à partir de cas, à savoir la phase d'adaptation avec notre contribution consistant à une intégration d'un autre type de raisonnement dans le cycle de RESPIDIAG.

Chapitre 5

La phase de recherche de RESPIDIAG

Sommaire

5.1 Introduction

La première étape du processus de RESPIDIAG est la phase de recherche. Elle consiste à mesurer la similarité entre les problèmes passés `srce` et le nouveau problème `tgt`, l'objectif étant de remémorer le cas $(\texttt{srce}, \texttt{sol}(\texttt{srce}))$ le plus proche du problème cible `tgt`.

Dans ce chapitre, nous détaillons la méthode avec laquelle RESPIDIAG sélectionne le cas le plus similaire, et nous exposons le problème du missing data rencontré lors de la réalisation de cette phase de recherche, empêchant ainsi son bon déroulement. Plusieurs approches sont proposées pour y remédier.

5.2 Les métriques de similarité

Nous avons adopté la méthode des k-plus proches voisins avec k=1, laissant les autres valeurs possibles pour le paramètre k pour des travaux ultérieurs. La similarité entre le problème cible et le cas source est estimée en deux étapes : d'abords, des similarités locales doivent être mesurées. Ces similarités sont calculées entre chaque deux valeurs du même attribut chez le problème cible `tgt` et le cas source `srce`. Par la suite, une agrégation est appliquée sur toutes les similarités locales, pour obtenir une similarité globale entre les deux cas en question.

L'équation suivante est utilisée pour évaluer la similarité globale entre `srce` et `tgt`.

$$\mathcal{S}(\texttt{srce}, \texttt{tgt}) = \frac{\sum\limits_{a \in \texttt{Attributs}} w_a \times \mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a)}{\sum\limits_{a \in \texttt{Attributs}} w_a} \tag{5.1}$$

où $w_a > 0$ est le poids de l'attribut a, (voir la section 5.2.1 pour plus de détails) et $\mathcal{S}_a$ est la mesure de similarité définie sur la plage de données de a.

5.2.1 Les coefficients de pondération

Après discussion avec les médecins, il s'est avéré que les dix sept attributs descripteurs considérés ne participent pas tous avec la même importance dans le diagnostic de la BPCO dans son état de base, ou dans ses exacerbations. Cette importance est reflétée dans RESPIDIAG par des coefficients de pondération représentant les poids de participation de chaque attribut dans la prise de décision.

Un travail avec nos médecins experts a permis de dresser les poids de ces attributs descripteurs en fonction de leurs importances dans le diagnostic. Il est à noter que ces poids ne dépendent pas seulement de l'attribut mais aussi de la classe de diagnostic associé à `srce`, i.e., $\texttt{sol}(\texttt{srce})$.

Cependant, et pour des raisons de lisibilité, nous avons choisi de ne pas le tenir en compte dans la notation : nous écrivons w_a au lieu de dire $w_a^{\texttt{sol(srce)}}$.

Par exemple, $w_{\texttt{toux}}^{\texttt{ExInfec}}$, le poids de l'attribut `toux` quand le diagnostic associé à `srce` est `ExInfec`, est simplement noté par $w_{\texttt{toux}}$ quand le contexte n'est pas ambigue.

La Table 5.1 donne les coefficients de pondération de chaque attribut en fonction des 6 classes de diagnostics. Nos médecins experts ont utilisé les notations suivantes : TGI pour désigner une Très Grande Importance, GI pour désigner une Grande Importance, MI pour désigner une Moyenne Importance et FI pour désigner une Faible Importance.

Par la suite nous avons attribué à chaque notation une valeur chiffrée.

Le poids w_a quand la classe de diagnostic de `srce` est $c = \texttt{sol}(\texttt{srce})$, figure sur la ligne de a et la colonne de c. Par exemple, si $c = \texttt{Stade\_II}$, $w_{\texttt{clarté}} = 0.4$.

		Stade_k		ExInfec		ExPnThor	
	Attribut	Import.	Coef	Import.	Coef	Import.	Coef
1	age	GI	0.6	GI	0.6	GI	0.6
2	antécédents	MI	0.4	MI	0.4	MI	0.4
3	profession	MI	0.4	MI	0.4	MI	0.4
4	température	FI	0.2	FI	0.2	FI	0.2
5	hyperLeucocytose	FI	0.2	TGI	0.8	FI	0.2
6	dyspnée	TGI	0.8	GI	0.6	TGI	0.8
7	toux	TGI	0.8	TGI	0.8	TGI	0.8
8	VEMS	TGI	0.8	GI	0.6	0	0
9	spirometrie	TGI	0.8	GI	0.6	0	0
10	clarté	MI	0.4	FI	0.2	TGI	0.8
11	opacité	MI	0.4	MI	0.4	FI	0.2
12	douleurThoracique	FI	0.2	FI	0.2	TGI	0.8
13	distensionThoracique	TGI	0.8	GI	0.6	TGI	0.8
14	râlesSibilants	GI	0.6	GI	0.6	GI	0.6
15	DEP	GI	0.6	GI	0.6	FI	0.2
16	tabagisme	TGI	0.8	GI	0.6	TGI	0.8
17	nombrePaquetParAnnée	TGI	0.8	GI	0.6	TGI	0.8

TABLE 5.1 – Les coefficients de pondération des attributs

5.2.2 Métriques de similarités des attributs numériques et booleens

L'estimation de la similarité locale $\mathcal{S}_a$ dépend du type de la plage des données de a. Lorsque ce type est booleen, la similarité est définie par :

$$\mathcal{S}_a(x,y) = \begin{cases} 1 & \text{si } x = y \\ 0 & \text{sinon} \end{cases} \qquad \text{pour } x, y \in \{\texttt{faux}, \texttt{vrai}\} \tag{5.2}$$

Si l'attribut a est de type numérique, $\mathcal{S}_a$ est défini par

$$\mathcal{S}_a(x,y) = 1 - \frac{|y-x|}{B_a} \tag{5.3}$$

où B_a est l'écart maximum de la plage de données de a, i.e., c'est l'écart (différence) entre la valeur maximale de la plage et sa valeur minimale. Cela peut être fait pour RESPIDIAG du moment que la plage de chaque attribut numérique dans cette application est bornée.

5.2.3 Métriques des similarités des attributs symboliques

Dans cette application de diagnostic de *BPCO*, nous avons six attributs parmi les dix sept de nature symbolique. Chacun possède une plage de données allant de quatre à vingts valeurs possibles. Généralement lorsqu'un attribut a est de type symbolique, la similarité entre deux valeurs x et y de a est estimée par la fonction binaire mentionnée dans la formule (5.2).

Dans un premier temps, cette formule a été utilisée pour nos attributs symboliques dans nos différentes approches proposées, et des résultats de cette utilisation sont fournis dans les sections 8.2, 8.3 et 8.4.

Par la suite, et dans l'objectif d'améliorer la qualité des résultats une "heuristique" a été proposée. Son principe est que la similarité entre deux valeurs x et y d'un attribut symbolique a, tel que $a.x \neq a.y$ ne doit pas être considérée toujours nulle. En effet, ces

deux valeurs $a.x$ et $a.y$ peuvent avoir des effets plus ou moins semblables sur la prise de décision d'un diagnostic donné, et donc leur similitude devra être entre 0 et 1 (graduelle). Il est à noter que suivant ce principe la similarité $\mathcal{S}_a(x, y)$ dépend aussi du diagnostic vers lequel nous voulons aller. Cependant, nous avons décidé de ne pas considérer cette dernière dépendance car elle est très coûteuse en terme de temps des experts. et nous avons travaillé de manière un peu générale pour estimer ces $\mathcal{S}_a(x, y)$.

En effet, des séances de travail avec nos médecins experts nous ont permis de définir pour chaque attribut symbolique, une table fixant la similarité entre chaque paire de valeurs appartenant à la plage de données de l'attribut concerné. Bien évidemment, les similarités sont estimées en fonction de leur impact sur le diagnostic de la *BPCO* de manière générale.

A titre d'exemple, la valeur "productive avec expectoration muqueuse" de l'attribut `toux` est estimée être similaire à la valeur "productive avec expectoration mucopurulente" à un taux de 80% alors qu'elle est estimée similaire à la valeur "sèche" juste à un taux de 20%

La similarité correspondante aux valeurs de l'attribut `profession` est estimée en fonction du risque d'avoir une *BPCO* causée par un milieu professionnel pollué. Donc pour deux milieux jugés pollués la similarité entre eux est 1, alors qu'entre un milieu professionnel sain et un autre pollué la similarité est considérée *nulle*.

Les tables 5.2, 5.3, 5.4 et 5.5 montrent les valeurs de similarités que nous avons dressées avec l'aide des médecins, entre chaque paire de valeurs des attributs `toux`, `spirometrie`, `dyspnée` et `tabagisme` respectivement.

$\mathcal{S}_{\texttt{toux}}(x, y)$ est le nombre se trouvant à la ligne de la valeur x et de la colonne de la valeur y. A titre d'exemple, $\mathcal{S}_{\texttt{toux}}(\texttt{PMP}, \texttt{PP}) = 0.8$, et $\mathcal{S}_{\texttt{spirometrie}}(\texttt{TVOR}, \texttt{TVOI}) = 0.4$.

Finalement, nous pouvons noter que pour chaque $a \in \texttt{Attributs}$ et pour chaque x et y appartenant à la plage de données de a, $\mathcal{S}(x, y) = 1$ si et seulement si $x = y$, mis à part pour l'attribut profession.

	pasDeToux	sèche	PM	PP	PMP
pasDeToux	1	0	0	0	0
sèche	0	1	0.2	0.2	0.2
PM	0	0.2	1	0.4	0.8
PP	0	0.2	0.4	1	0.8
PMP	0	0.2	0.8	0.8	1

TABLE 5.2 – Similarités entre paires de valeurs de l'attribut `toux`.

	normal	TVR	TVOI	TVOR
normal	1	0	0	0
TVR	0	1	0	0
TVOI	0	0	1	0.4
TVOR	0	0	0.4	1

TABLE 5.3 – Similarités entre paires de valeurs de l'attribut `spirometrie`.

	stade 0	stade 1	stade 2	stade 3	stade 4
stade 0	1	0	0	0	0
stade 1	0	1	0.75	0.50	0.25
stade 2	0	0.75	1	0.75	0.50
stade 3	0	0.50	0.75	1	0.75
stade 4	0	0.25	0.50	0.75	1

TABLE 5.4 – Similarités entre paires de valeurs de l'attribut `dyspnée`.

	$Absent$	$Passif$	$Actif$
$Absent$	1	0	0
$Passif$	0	1	0.75
$Actif$	0	0.75	1

TABLE 5.5 – Similarités entre paires de valeurs de l'attribut `tabagisme`.

5.3 Le problème des données manquantes

Parfois le médecin est en position de prise de décision à propos d'un diagnostic pour lequel quelques données sont absentes. C'est le cas par exemple dans les urgences lorsque le patient arrive à l'hopital dans un état de crise respiratoire, alors le médecin est en obligation de faire un diagnostic et de commencer une prise en charge immédiate sans avoir le temps des fois de demander toutes les informations nécessaires ou d'attendre les résultats d'examinations indispensables (radiologie, analyses, ...) à la prise de décision. Parfois aussi le médecin peut décider d'éviter certains examens qui peuvent s'avèrer trop coûteux ou plutôt trop risqués pour le patient, et donc il se charge du diagnostic malgré les données manquantes. Les patients présentant une *BPCO* arrivent souvent dans un état de crise au service de pneumologie, ce qui entraîne des données manquantes dans leurs dossiers.

Le problème du missing data soulève une difficulté sérieuse de concevoir la procédure de recherche de RESPIDIAG. En effet, la valeur $\mathcal{S}(\texttt{srce}, \texttt{tgt})$ ne peut être calculée si au moins pour un attribut a, $\texttt{srce}.a$ ou $\texttt{tgt}.a$ est inconnue. Et si cette similarité ne peut être calculée, la phase de recherche ne peut être achevée et donc le processus *RàPC* ne peut avoir lieu !

Dans ce qui suit, la notation "$\texttt{pb}.a = \texttt{?}$" (resp., "$\texttt{pb}.a \neq \texttt{?}$") signifie que la valeur de a pour le problème `pb` est inconnue (resp., connue).

La manière dont nous traitons le problème de ces données manquantes est l'une des contributions de cette thèse.

5.4 Description de la base de cas

Notre base de cas est composée de dix sept attributs correspondant aux descripteurs du problème déjà mentionnés dans la section 4.3, en plus d'un autre attribut qui correspond à la solution du problème : le diagnostic. Avec la collaboration des médecins, 40 cas réels ont été collectés de l'archive du service de pneumologie de l'hôpital Dorban (Annaba). Cet ensemble de cas est considéré par les experts comme étant assez suffisant pour une représentation initiale de nos six diagnostics et il a été selectionné de facon à avoir un maximum de diversité dans les symptômes pour le même diagnostic.

Pour les différentes raisons exposées dans la section 5.3, notre base de cas collectée à partir des dossiers de patients, contient des données manquantes pouvant apparaitre dans différents attributs des cas, et dont le taux atteint les 21.61% de l'ensemble des données. Il est à noter également que dans le même cas `srce` ou `tgt`, il peut y avoir plusieurs attributs non renseignés. La table 5.6 montre quelques statistiques sur notre base de cas.

	Nombre	Pourcentage
Cas	40	
Données présentes	533	78.39 %
Données manquantes	147	21.61 %
Moyenne des données absentes par cas	3.67	21.58 %

TABLE 5.6 – Statistiques sur la base de cas

Soient `srce` et `tgt` les deux problèmes à comparer et a, un attribut pour lequel la valeur est inconnue dans `srce` ou dans `tgt`, ou dans les deux en même temps :

$\texttt{srce}.a = \texttt{?}$ et/ou $\texttt{tgt}.a = \texttt{?}$.

Notre contribution pour la phase de recherche de RESPIDIAG est basée sur deux questions principales :

- "Qu'est ce qui pourrait être utilisé pour la valeur de $\mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a)$ pour achever le calcul de $\mathcal{S}(\texttt{srce}, \texttt{tgt})$ en fonction de l'équation (5.1) ?"
- "Qu'est ce qui pourrait être utilisé pour la valeur manquante -elle même- de a pour achever le calcul de $\mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a)$?"

Dans cette thèse, nous avons répondu à ces questions par la proposition et l'évaluation de plusieurs approches qui traitent toutes ce problème de missing data mais en ayant des principes différents et même des objectifs différents. Dans la pratique, ces approches n'ont pas été proposées ni évaluées toutes en même temps. En effet, un décalage de plusieurs mois a eu lieu entre la proposition et l'évaluation des trois premières approches, de celles qui ont suivi par la suite. Pour cette raison, nous les présentons séparément, même si quelques unes des deux parties sont liées.

La section 5.5 présente les trois premières approches proposées qui seront évaluées et discutées dans le chapitre 8, alors que la section 5.6 détaille les autres stratégies dont les évaluations sont exposées dans le chapitre 9 de cette thèse.

5.5 Les premières approches proposées

Dans cette section nous proposons trois approches pour gérer le problème du missing data. Ces approches utilisent chacune un principe différent mais visent toutes à attribuer une valeur à la similarité locale entre deux attributs, à chaque fois que l'une des deux valeurs (ou les deux en même temps) est (sont) absente(s). Ces trois stratégies appelées *pessimiste*, *médium* et *sélective* [Guessoum12-b, Guessoum12-c] affectent alors des différentes valeurs aux similarités pendant le processus *online* de RESPIDIAG, plus préci-

sément pendant sa phase de recherche. Une variante de l'approche *pessimiste* que nous avons appelée *pessimiste*** sera présentée aussi.

5.5.1 L'approche *pessimiste* et sa variante *pessimiste***

Dans cette approche, et partant d'un point de vue assez pessimiste, nous proposons de mettre en évidence la possibilité que la valeur manquante soit la plus loin possible de la valeur présente. Ce principe pessimiste fait que la similarité locale entre les deux valeurs va être considérée *nulle* à chaque fois que l'une des valeurs (ou les deux en même temps) est (sont) absente(s).

En supposant que a est l'attribut concerné par le calcul de similarité locale, si nous avons : $\texttt{tgt}.a = \texttt{?}$ et/ou $\texttt{srce}.a = \texttt{?}$, alors nous obtiendrons :

$$\mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a) := 0 \qquad \text{(stratégie pessimiste)}$$

Dans cette approche, nous continuons à considérer dans l'équation globale (5.1), les valeurs des coefficients de pondération des attributs w_a tel que :

$\texttt{tgt}.a = \texttt{?}$ et/ou $\texttt{srce}.a = \texttt{?}$

Dans une première variante de cette approche *pessimiste*, les similarités locales entre les valeurs présentes des attributs de type symbolique sont considérées selon l'équation (5.2). Dans une deuxième variante, nous considérons les "fonctions heuristiques" proposées dans la section 5.2.3 pour les attributs symboliques. Cette deuxième variante et pour des raisons de distinction sera appelée au niveau des expérimentations *pessimiste***.

5.5.2 L'approche *médium*

Dans l'approche médium, nous avons une vision un peu plus optimiste que celle de l'approche précédente. Nous mettons en évidence la possibilité que la valeur manquante peut être moyennement près/loin de la valeur présente. Et même si elles sont toutes les

deux absentes, nous supposons que l'une est moyennement près/loin de l'autre. Nous attribuons la valeur moyenne de l'intervalle [0,1] à la similarité cherchée. Donc, pour un attribut a tel que $\texttt{tgt}.a = \texttt{?}$ et/ou $\texttt{srce}.a = \texttt{?}$, nous mettons

$$\mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a) := 0.5 \qquad \text{(stratégie médium)}$$

Au niveau des expérimentations, cette approche a été testée en considérant les similarités binaires pour les attributs de type symbolique lorsque nous avons en même temps : $\texttt{tgt}.a \neq \texttt{?}$ et $\texttt{srce}.a \neq \texttt{?}$

5.5.3 L'approche *sélective*

Dans cette approche, nous proposons qu'à chaque comparaison entre le problème cible et un cas source, une sélection se fait sur les attributs renseignés dans cible et source en même temps. Les attributs où au moins $\texttt{tgt}.a = \texttt{?}$ ou $\texttt{srce}.a = \texttt{?}$ sont ignorés. Le principe est donc de raisonner juste sur ce qui est connu. Les poids des attributs non renseignés ne sont pas comptabilisés dans le total des poids de l'agrégation de l'équation (5.1) qui devient alors :

$$\mathcal{S}(\texttt{srce}, \texttt{tgt}) = \frac{\sum\limits_{\substack{a \in \texttt{Attributs} \\ \texttt{srce}.a \neq \texttt{?} and \texttt{tgt}.a \neq \texttt{?}}} w_a \times \mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a)}{\sum\limits_{a \in \texttt{Attributs}} w_a} \tag{5.4}$$

Les expérimentations et l'évaluation de ces approches qui viennent d'être présentées : *pessimiste*, *pessimiste***, *médium* et *sélective* sont exposées dans le chapitre 8.

Dans la section suivante, nous apportons une amélioration dans le principe des deux approches *pessimiste* et *médium*, puis nous proposons une nouvelle approche *optimiste* qui s'exécute aussi à l'état *online* de RESPIDIAG.

Nous présentons également dans cette section deux autres approches qui traitent toujours le problème des données manquantes mais diffèrent dans leurs objectifs des autres stratégies. Elles s'exécutent à l'état *offline* de RESPIDIAG.

5.6 Les deuxièmes approches proposées

Comme nous venons de le mentionner à la fin de la section précédente, nous présentons ici trois différentes approches *online* pour estimer la similarité $\mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a)$ malgré les données manquantes. Ces approches se distinguent par leur manière de considérer la valeur absente dans le calcul des similarités locales. Les deux premières ne sont que de nouvelles variantes des approches *online pessimiste* et *médium* de la section précédente mais qui voient leur principes améliorés, elles sont nommées *pessimiste** et *médium** [Guessoum13]. La troisième est l'approche *optimiste* [Guessoum13].

Nous proposons également dans cette section deux autres approches *offline* consistant à combler les vides laissés par les données manquantes dans la base de cas.

5.6.1 L'approche *pessimiste* online*

Toujours pendant le processus *online* de RESPIDIAG et dans la même optique de la section 5.5.1, une stratégie pessimiste vis à vis de la valeur manquante peut supposer qu'elle est la plus loin possible de la valeur présente. Donc la contribution de la similarité locale $\mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a)$ à la similarité globale $\mathcal{S}(\texttt{srce}, \texttt{tgt})$ est minimale.

Dans la section 5.5.1, l'approche et par son principe pessimiste attribuait la valeur *nulle* à la similarité locale à chaque fois qu'il y a une valeur manquante dans l'attribut. En réalité, cette affectation est faite sans prise en compte du type de cet attribut.

Dans cette section la même approche considère la nature de l'attribut pour décider de la valeur la plus pessimiste à affecter à la similarité locale $\mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a)$. Cette

valeur dépend cette fois-ci et étroitement de la plage de données et du type de l'attribut en question.

Ce qui donne :

$$\text{if } \texttt{srce}.a = \texttt{?} \text{ and } \texttt{tgt}.a \neq \texttt{?} \quad \mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a) := \inf_{x \in \texttt{plage}(a)} \mathcal{S}_a(x, \texttt{tgt}.a)$$

$$\text{if } \texttt{srce}.a \neq \texttt{?} \text{ and } \texttt{tgt}.a = \texttt{?} \quad \mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a) := \inf_{y \in \texttt{plage}(a)} \mathcal{S}_a(\texttt{srce}.a, y)$$

$$\text{if } \texttt{srce}.a = \texttt{?} \text{ and } \texttt{tgt}.a = \texttt{?} \quad \mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a) := \inf_{x,y \in \texttt{plage}(a)} \mathcal{S}_a(x, y)$$

(Stratégie pessimiste)

A titre d'exemple, si $\texttt{srce.VEMS} = \texttt{?}$ et $\texttt{tgt.VEMS} = 40\%$ alors la valeur choisie par la stratégie pessimiste est $\mathcal{S}_{\texttt{VEMS}}(\texttt{srce.VEMS}, \texttt{tgt.VEMS})$ est $1 - \frac{100\% - 40\%}{100\%} = 0.4$, puisque la plage de données de `VEMS` est $[0\%, 100\%]$.

Lorsque l'absence est dans un attribut binaire ou symbolique, la valeur manquante est supposée opposée à la valeur présente et alors la similarité locale prendra tout simplement la valeur *nulle*, et là les deux variantes coincident. Par contre lorsque la valeur manquante est de type numérique, la distance est estimée en supposant que cette valeur est la plus loin possible dans la plage de donnée de l'attribut, et la similarité la plus pessimiste ne pourra pas être *forcément nulle* !

La table 5.7 donne un pseudo-code de cette approche.

if (BaseDeCas.$a.type$ = symbolic) or (BaseDeCas.$a.type = binary$) then
if (srce.a = ?) or (tgt.a = ?) then $\mathcal{S}_a$(srce.a, tgt.a) := 0
else if (BaseDeCas.$a.type$ = numeric) then
if (tgt.a = ?) and (srce.$a \neq$?) then
if (tgt.a > (BaseDeCas.$a.plage.middle$) then
tgt.a := BaseDeCas.$a.plage.minimum$
else tgt.a := BaseDeCas.$a.plage.maximum$

TABLE 5.7 – Pseudo-code pour estimer la similarité pessimiste

5.6.2 L'approche *optimiste online*

Dans cette politique, nous changeons de vision et nous partons d'un esprit très optimiste : la valeur manquante peut être très proche, la plus proche possible même de la valeur présente. Et même si les deux valeurs sont absentes, nous mettons en évidence la possibilité qu'elles peuvent coincider. Donc la contribution de la similarité locale $\mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a)$ à la similarité globale $\mathcal{S}(\texttt{srce}, \texttt{tgt})$ est maximale, ce qui donne :

$$\mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a) := 1 \qquad \text{(stratégie optimiste)}$$

Il est à noter que dans cette approche, il n y pas de différences dans la valeur affectée à la similarité locale en fonction du type de l'attribut concerné, ce qui était le cas dans la stratégie *pessimiste**. La table 5.8 donne un pseudo-code du principe de l'approche *optimiste*.

if ($\texttt{srce}.a = \texttt{?}$) or ($\texttt{tgt}.a = \texttt{?}$) then
$\mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a)$:= 1 *(indépendemment du type de l'attribut a)*

TABLE 5.8 – Pseudo-code pour estimer la similarité optimiste

5.6.3 L'approche *médium* online*

Dans la section 5.5.2, nous avons proposé une approche *médium* dont le principe est d'estimer moyennement la similarité locale lorsque l'une des deux valeurs concernées par le calcul est absente, ou même les deux. Cette estimation moyenne attribuait la valeur de 0.5 à la similarité locale.

Dans cette section, la même stratégie voit son principe s'améliorer pour donner une nouvelle variante que nous avons appelée *médium**. En effet, l'estimation moyenne ne donne plus une valeur fixée à 0.5 à la similarité locale mais plutôt une valeur moyenne calculée à la base des deux stratégies précédentes, à savoir la *pessimiste** et *l'optimiste*.

Au niveau de l'implémentation, l'utilisation de l'approche *médium** fait automatiquement appel juste à l'approche *pessimiste** pour avoir une valeur *pessimiste* notée par v_p dans ce qui suit. Pour la valeur *optimiste*, c'est directement la valeur 1 qui est utilisée dans la formule de cette stratégie *médium**.

$$\mathcal{S}_a(\texttt{srce}.a, \texttt{tgt}.a) := \frac{1 + v_p}{2} \qquad \text{(stratégie médium*)}$$

Avec l'exemple de la section 5.6.1, la valeur choisie par la stratégie médium pour $\mathcal{S}_{\texttt{VEMS}}(\texttt{srce.VEMS}, \texttt{tgt.VEMS})$ est $\frac{1+0.4}{2} = 0.7$.

Dans les deux sous sections suivantes nous proposons deux approches qui diffèrent dans leurs objectifs de celles que nous avons présentées jusqu'ici, malgré qu'elles visent

aussi le traitement du missing data. Elles consistent à un comblement des vides laissés par les données manquantes dans la base de cas, exploitant chacune un principe différent pour y arriver. Donc elles sont utilisées pour attribuer une valeur plausible v à $\texttt{srce}.a$, lorsque $\texttt{srce}.a = \texttt{?}$. Cette attribution est notée dans ce qui suit par $\texttt{srce}.a := v$.

Les deux stratégies qui suivent s'exécutent à l'état *offline* de RESPIDIAG, et concernent uniquement le manque de données dans la base de cas et non dans le problème cible.

5.6.4 L'approche statistique *offline*

La première de ces approches est basée sur une méthode statistique [Guessoum14]. Son principe consiste à combler le vide laissé par la valeur manquante dans un attribut a de la BaseDeCas, par la valeur moyenne des valeurs présentes dans le même attribut a.

Soit $\texttt{SPWKV}_a$ l'ensemble des problèmes sources avec valeurs connues de l'attribut a :

$$\texttt{SPWKV}_a = \{\texttt{srce}' \mid (\texttt{srce}', \texttt{sol}(\texttt{srce}')) \in \texttt{BaseDeCas} \text{ et } \texttt{srce}'.a \neq \texttt{?}\}$$

Deux situations doivent être considérées, en fonction du type de l'attribut a : numérique ou non.

Quand l'attribut a est numérique, nous proposons dans cette approche de combler la valeur manquante de l'attribut a dans srce par la moyenne des valeurs de a dans les problèmes sources pour lesquels ces valeurs sont connues (donc appartenant à $\texttt{SPWKV}_a$). Cela est traduit par l'équation suivante :

$$\texttt{srce}.a := \frac{\sum_{\texttt{srce}' \in \texttt{SPWKV}_a} \texttt{srce}'.a}{\text{card } \texttt{SPWKV}_a} \quad \text{(stratégie statistique pour les attributs numériques)}$$

où card X est les nombre des éléments de l'ensemble fini X.

Lorsque a est non numérique (i.e., booleen ou symbolique), nous proposons dans

cette approche statistique de procéder à un vote. La valeur manquante de l'attribut a sera la valeur dont le nombre d'occurences est maximum parmi tous les problèmes sources pour lesquels l'attribut a est renseigné.

$$\texttt{srce}.a := \underset{v \in \texttt{plage}(a)}{\text{argmax}} \ \text{card}\ \{\texttt{srce}' \in \texttt{SPWKV}_a \mid \texttt{srce}'.a = v\}$$

(stratégie statistique pour les attributs non numériques)

Lorsqu'il apparait plusieurs valeurs avec le même nombre maximal d'occurences dans la base de cas, d'autres stratégies peuvent être utilisées pour trancher dans le choix, par exemple prendre la première valeur dont le nombre d'occurences est maximal.

5.6.5 L'approche RàPC *offline*

Dans cette approche, et toujours dans l'objectif de combler la valeur manquante $\texttt{srce}.a$ par une valeur plausible, nous nous proposons cette fois-ci d'exploiter le principe du RàPC lui-même.

Ce processus RàPC utilisé pour proposer une valeur à $\texttt{srce}.a$, sera appelé dans ce qui suit : *"processus RàPC pour l'estimation de a"* [Guessoum14], à l'opposé de *"processus RàPC pour le diagnostic de la BPCO"*, qui était le but principal de RESPIDIAG.

Cela signifie que dans ce nouveau *"processus RàPC pour l'estimation de a"*, le cas source est décomposé d'une autre façon que précédemment. Un problème est défini par l'ensemble de tous les attributs de $\texttt{srce}$ excepté a et la solution est une valeur pour a.

Par conséquent, un cas source pour cette approche est un problème source $\texttt{srce}$ pour le *"processus RàPC pour le diagnostic de la BPCO"*.[2]

Un cas cible dans cette approche est un problème source $\texttt{srce}$ pour le *"processus RàPC pour le diagnostic de la BPCO"* contenant un attribut a non renseigné. Ce qui donne en fait que le travail qui suit va porter sur les problèmes sources de notre base de

2. Il est possible aussi de prendre en compte la solution $\texttt{sol}(\texttt{srce})$ comme un attribut du *problème*. Cependant, nous n'allons pas le considérer dans ce qui suit, pour des raisons de lisibilité.

cas.

Le fait que les définitions des cas cible et source ont changé relativement à notre premier processus RàPC, une nouvelle mesure de similarité $\mathcal{S}^a$ doit être définie entre ces nouveaux cas sources, prenant en considération uniquement la partie problème. Nous proposons d'utiliser l'équation (5.5).

$$\mathcal{S}^a(\texttt{srce}_1, \texttt{srce}_2) = \frac{\sum_{b \in \texttt{Attributs} \setminus \{a\}} w_b^a \times \mathcal{S}_b(\texttt{srce}_1.b, \texttt{srce}_2.b)}{\sum_{b \in \texttt{Attributs} \setminus \{a\}} w_b^a} \tag{5.5}$$

w_b^a est le poids de l'attribut b pour le *"processus RàPC pour l'estimation de a"*.

Notons que le poids w_b^a peut être différent du poids w_b utilisé pour le *"processus RàPC pour le diagnostic de la BPCO"* et des poids w_b^c pour le *"processus RàPC d'estimation c"*, où c est un autre attribut. [3]

Un autre problème peut être rencontré pendant le *"processus RàPC pour l'estimation de a"* : Il peut apparaitre que $\texttt{srce}_1.b = \texttt{?}$ et/ou $\texttt{srce}_2.b = \texttt{?}$, tel que $a \neq b$.

Dans ce cas, nous proposons d'exploiter l'une des approches *online* présentées dans les sections 5.6.2, 5.6.1 et 5.6.3.

Dans l'implémentation de RESPIDIAG, nous avons opté pour l'utilisation de l'approche *pessimiste**, mais nous aurrions pu aussi utiliser l'approche *optimiste* ou *médium* ici, ces options peuvent être étudiées dans des travaux ultérieurs.

Ainsi, une solution serait de réutiliser la valeur $\texttt{srce}'.a$ du problème source le plus proche à $\texttt{srce}$, selon $\mathcal{S}^a$.

Cependant, nous avons choisi de prendre en compte *tous* les cas sources $\texttt{srce}' \in \texttt{SPWKV}_a$, mais en donnant plus d'importance à ceux similaires à $\texttt{srce}$ selon $\mathcal{S}^a$.

3. Nous aurions pu définir de nouvelles mesures de similarité locale $\mathcal{S}_b$, mais nous nous avons choisi de les conserver.

Cette approche RàPC *offline* que nous proposons est partiellement similaire à l'approche statistique que nous avons présentée dans la section 5.6.4. La différence majeure étant que dans la deuxième, les valeurs $\texttt{srce}'.a$ des cas sources $\texttt{srce}' \in \texttt{SPWKV}_a$ sont pondérées chacune par la similarité $\mathcal{S}^a(\texttt{srce}', \texttt{srce})$.

Cela signifie que si a est un attribut non renseigné de `srce`, alors :

$$\texttt{srce}.a := \frac{\sum_{\texttt{srce}' \in \texttt{SPWKV}_a} \mathcal{S}^a(\texttt{srce}', \texttt{srce}) \times \texttt{srce}'.a}{\sum_{\texttt{srce}' \in \texttt{SPWKV}_a} \mathcal{S}^a(\texttt{srce}', \texttt{srce})}$$

(Stratégie RàPC pour les attributs numériques)

La formule précédente ne peut en réalité être utilisée que pour les attributs de type numérique. Pour les attributs de type symbolique ou binaire, nous nous proposons d'exploiter le principe du vote pondéré par les similarités. Le vote consiste à estimer la valeur ayant le plus élevé nombre d'occurences pondéré par la similarité.

$$\texttt{srce}.a := \underset{v \in \texttt{plage}(a)}{\text{argmax}} \sum \{\mathcal{S}^a(\texttt{srce}', \texttt{srce}) \mid \texttt{srce}' \in \texttt{SPWKV}_a \text{ and } \texttt{srce}'.a = v\}$$

(Stratégie RàPC pour les attributs non numériques)

où $\sum X = \sum_{x \in X} x$ pour tout ensemble fini de nombres X.

5.7 Conclusion

Dans ce chapitre, nous avons proposé plusieurs approches visant le traitement du missing data, aussi bien dans le problème *cible* que dans les cas *sources*. Quatre parmi elles (avec sept variantes en tout), consistent à attribuer une valeur à la similarité locale pendant le processus de recherche de RESPIDIAG, et ce à chaque fois que l'une des deux valeurs concernées par la similarité est absente. Les deux dernières approches visent un

autre objectif, combler la valeur absente par la valeur la plus plausible selon un principe bien déterminé. Elles s'exécutent pendant l'état offline de RESPIDIAG et ne concernent que les données manquantes dans la base de cas. La table 5.9 donne une synthèse des différentes approches proposées.

Le chapitre suivant est consacré à la phase d'adaptation de RESPIDIAG. Il détaille notre contribution pour cette phase consistant à faire une intégration d'un deuxième mode de raisonnement à l'intérieur du processus de RESPIDIAG.

N	Approche	Objectif : affecter une valeur	La valeur affectée	Etat	Sim. pour les attributs symboliques / Observation
1	*pessimiste*	à la similarité locale	la valeur 0	*online*	binaire
2	*médium*	à la similarité locale	la valeur 0.5	*online*	binaire
3	*sélective*	sélectionner les attributs renseignés		*online*	binaire
4	*pessimiste***	à la similarité locale	la valeur 0	*online*	graduelle
5	*pessimiste**	à la similarité locale	la valeur la + faible possible	*online*	graduelle
6	*médium**	à la similarité locale	la moyenne de deux valeurs	*online*	graduelle
7	*optimiste*	à la similarité locale	la valeur 1	*online*	graduelle
8	*statistique*	à la donnée manquante.	moyenne des valeurs présentes/ vote	*offline*	il faut une appr. online pour avoir le diagnostic
9	*RàPC*	à la donnée manquante	moyenne des val. présentes pondérées par la sim./ vote pondré	*offline*	appelle une app. online pendant son propre processus,

TABLE 5.9 – Récapitulatif des approches proposées pour la phase de recherche

Chapitre 6

La phase d'adaptation de RESPIDIAG

Sommaire

6.1 Introduction

Le principe de la phase d'adaptation est d'apporter des modifications sur la solution du cas source retenu de la phase précédente pour la rendre adaptée aux données du problème cible. Cette phase est la plus délicate à concevoir dans tout le processus RàPC. De multiples travaux de recherches se sont penchés sur cette phase dans l'objectif de réduire sa complexité, mais jusqu'à l'heure actuelle sa difficulté résiste encore. Cela relève principalement du fait que cette phase est en dépendence étroite et totale des connaissances du domaine d'application. Et partant d'une application à une autre, les choses ne peuvent

être vues du même angle, et par conséquent la manière de concevoir cette phase change totalement. Dans le domaine médical, cette phase prend encore plus de délicatesse. Nous n'ouvrons pas droit à l'erreur ! Le diagnostic doit être le plus précis possible, et le traitement doit être aussi le plus efficace possible.

Une collaboration avec les experts du domaine était alors indispensable durant toute la conception de cette phase pour dresser des mécanismes qui permettront de passer de la solution du cas source retenu vers la solution à proposer pour le cas cible.

6.2 La phase d'adaptation de RESPIDIAG

Dans la section 3.3 nous avons donné un aperçu sur l'apport des systèmes experts à base de règles dans la médecine, et nous avons montré leur succès à travers les nombreux systèmes médicaux qui ont été développés pour le diagnostic ou/et la thérapie de certaines pathologies.

Voulant exploiter la performance des systèmes experts (SE) dans le raisonnement à base de règles, nous proposons dans ce chapitre une conception de SE pour modéliser le processus d'adaptation de RESPIDIAG. Cette proposition est l'une des contributions de cette thèse [Guessoum06, Guessoum07].

L'architecture du système expert qui va être intégré aura alors une base de connaissances et un moteur d'inférences. La base de connaissances étant composée d'une base de faits et d'une base de règles qui constitue le coeur de conception dans un système expert. Pour RESPIDIAG, le processus SE sera appelé à la fin du processus de recherche, en lui transmettant les données nécessaires qui constitueront les faits initiaux à ses inférences. Ces données transmises seront détaillées dans la section 6.2.2. Notons que nous avons opté pour une stratégie de chaînage avant pour le fonctionnement de notre système expert. L'essai des deux autres chaînages (arrière et mixte) est laissé pour des travaux ultérieurs.

6.2.1 La base de règles du système expert

Avec la collaboration de nos médecins experts, nous avons établi un ensemble de règles (environ une trentaine) visant l'adaptation du diagnostic du cas source retenu $(\texttt{srce}, \texttt{sol}(\texttt{srce}))$ au problème cible `tgt`. Cet ensemble de règles constitue la base de règles de notre système expert. Notons que le développement de ces règles a été une tâche assez compliquée pour nous puisque la connaissance de l'adaptation était difficile à exprimer par les experts et donc pour nous, difficile à modéliser.

Etant un diagnostic du cas source retenu $\texttt{sol}(\texttt{srce})$, nous avons basé les prémisses de nos règles -partant de ce diagnostic- sur les attributs ayant les poids les plus élevés.

Comme indiqué dans la section 4.1, les diagnostics de notre champs d'étude sont :

- $BPCO$ `Stade_`k, avec $k \in (I, II, III, IV)$,
- $Ex.Infec$, pour "Exacerbation de BPCO d'origine infectieuse"
- $Ex.Pn.Th$ pour "Exacerbation de BPCO d'origine Pneumo-Thorax".

Pour modéliser la phase d'adaptation, nous avons décomposé les désignations de ces diagnostics en deux parties comme le montre la table suivante :

Partie 1	Partie 2
BPCO	$Stadek$
Exacerbation de BPCO	d'origine infectieuse
Exacerbation de BPCO	d'origine Pneumo-thorax

TABLE 6.1 – Les deux parties de nos diagnostics

Les règles d'adaptation que nous avons établies pour le système expert de RESPIDIAG portent en réalité sur la seconde partie du diagnostic. La première ne subira pas de changements dans cette phase et le système va la garder telle qu'elle a été trouvée dans le cas source retenu par la phase de recherche. Cela signifie que nous pourrons passer entre les différents stades de la $BPCO$, ou entre les deux exacerbations. Le passage entre une

$BPCO$ `Stade_`k et les exacerbations n'étant pas possible du moins dans ce travail. La raison principale étant la complication de la modélisation de ce passage par des règles que nous essayerons de surmonter dans des travaux ultérieurs.

A titre d'exemple, pour un diagnostic d'un cas `srce` récupéré de la phase de recherche, $BPCO\ stadeII$ nous retenons la première partie $BPCO$ sans changement et nous substituons $stadeII$ par exemple par $stadeIV$ en fonction des attributs du problème `tgt`.

Un échantillon de règles d'adaptation est donné dans la table 6.2.

R1	**if** `sol(srce)` $=$ `Stade_I` **and** `tgt.VEMS` $\leq 30\%$ **then** `sol(tgt)= Stade_IV`
R2	**if** `sol(srce)` $=$ `Stade_II` **and** `tgt.VEMS` $\leq 50\%$ **and** `tgt.VEMS` $\succcurlyeq 30$ **then** `sol(tgt)= Stade_III`
R3	**if** `sol(srce)` $=$ `ExPnThor` **and** `tgt.hyperLeucocytose` $= yes$ **then** `sol(tgt)= ExInfec`
R4	**if** `sol(srce)` $=$ `ExInfec` **and** `tgt.douleurThoracique` $= yes$ **then** `sol(tgt)= ExPnThor`

TABLE 6.2 – Un échantillon de règles d'adaptation

6.2.2 La base de faits du système expert

A la fin de la phase de recherche, le système expert doit entamer la tâche d'adaptation. Pour ce fait, il a besoin de certaines informations qui lui seront transmises en tant que des faits initiaux nécessaires à ses inférences. Le premier fait qui lui est transmis est bien le diagnostic récupéré du cas `srce` retenu de la phase de recherche, et à partir duquel une sélection est faite sur les valeurs des attributs ayant les coefficients les plus forts

relativement à ce diagnostic. Ces valeurs sélectionnées sont transmises à la base de faits du système expert. Ainsi, le moteur d'inférences est déclenché sur cet ensemble de faits selon une stratégie de chaînage avant.

6.3 Les phases de révision et d'apprentissage

Les deux dernières phases du cycle RàPC sont bien la révision et l'apprentissage. Dans RESPIDIAG ces phases ne font pas l'objet d'une étude approfondie ou d'une quelconque contribution. Elles sont réalisées dans RESPIDIAG de la manière la plus simple et la plus évidente, laissant son approfondissement pour un travail ultérieur.

En effet, une fois le diagnostic est adapté par le système expert, il est présenté à l'utilisateur qui procèdera à la validation, l'interface de RESPIDIAG lui permettant la confirmation ou la modification du diagnostic en cas de besoin (échec du système). Par la suite, RESPIDIAG procède à l'ajout du nouveau problème avec sa solution (ce qui constitue un nouveau cas appris pour le système) dans la base de cas.

6.4 Conclusion

Dans cette section, nous avons détaillé le processus d'adaptation de RESPIDIAG qui est modélisé par un système expert fonctionnant en stratégie de chaînage avant. Environ une trentaine de règles ont été établies avec la collaboration de nos médecins spécialistes. Leur objectif est d'apporter une adaptation adéquate au diagnostic récupéré du cas `srce` retenu dans la phase de recherche pour que la solution proposée par RESPIDIAG soit conforme au cas du nouveau problème (patient). Un échantillon de ces règles a été donné dans cette section suivi d'une présentation des données transmises au système expert.

Ce chapitre mentionne par ailleurs que les deux dernières phases du cycle RàPC de RESPIDIAG sont réalisées de la manière la plus évidente.

Ainsi se termine notre présentation détaillée du cycle de RESPIDIAG et nous entamons dans la troisième partie de cette thèse, les différentes expérimentations réalisées ainsi que les résultats obtenus.

Troisième partie

Implémentation et évaluations

Chapitre 7

Introduction

Sommaire

7.1 Les données et la méthode d'évaluation

Dans cette troisième partie, nous présentons l'implémentation de RESPIDIAG, ainsi que les détails de l'évaluation des différentes propositions de cette thèse. Nous travaillons sur la base de cas que nous avons décrite déjà dans la section 5.4. Nous pouvons rappeler son état dans la table 7.1 qui montre quelques statistiques sur les données présentes/absentes. Les cas de cette base sont pris des dossiers patients du service de pneumologie.

L'échantillon de test a été sélectionné aussi de la même source grâce à la collaboration de nos médecins experts. Il consiste en 21 cas réels avec quelques données manquantes. La table 7.2 montre quelques statistiques sur l'état de cet échantillon de test.

La méthode d'évaluation consiste en la saisie des symptômes d'un cas de test qui représente le problème `tgt` pour RESPIDIAG, lancer le processus de la phase de recherche

et récupérer son résultat, avant même le déclenchement de la phase d'adaptation. Cette étape va nous permettre de tester et de comparer les différentes approches proposées au profit du processus de recherche. Par la suite, nous récupérons les résultats de la phase d'adaptation pour compléter le travail de l'évaluation du système expert de RESPIDIAG.

	Nombre	Pourcentage
Cas	40	
Données absentes	533	78.39 %
Données manquantes	147	21.61 %
Moyenne des données absentes par cas	3.67	21.58 %

TABLE 7.1 – Statistiques sur la base de cas

	Nombre	Pourcentage
Cas	21	
Total des données	357	100 %
Données présentes	306	85.72 %
Données absentes	51	14.28 %
Moyenne des données absentes par cas	2.42	11.56 %

TABLE 7.2 – Statistiques sur l'échantillon de test.

Pour des raisons de lisibilité, nos classes de diagnostics seront représentées dans les tables de résultats par des chiffres allant de 1 à 5 comme le montre la table 7.3, étant que dans l'échantillon de test, nous n'avons pas eu de cas présentant le diagnostic de $BPCO$ `Stade_I`, donc nous avons évalué les approches de RESPIDIAG juste sur cinq classes de diagnostics.

Code	Diagnostic
1	BPCO stade II
2	BPCO stade III
3	BPCO stade IV
4	Exacerbation de BPCO d'origine infectieuse
5	Exacerbation de BPCO d'origine pneumo-thorax

TABLE 7.3 – Numérotation des classes de diagnostics

7.2 L'interface de RESPIDIAG

L'interface utilisateur de RESPIDIAG permet en premier lieu de faire entrer tous les symptômes observés chez le patient. Les valeurs saisies vont construire pour nous le nouveau problème à traiter. Pour simplifier l'entrée des symptômes, toutes les valeurs possibles pour un symptôme donné sont listées dessous, donc l'utilisateur clique simplement sur la valeur correspondante du patient. Le symbole "?" signifie que le médecin ignore la valeur du symptôme, ce qui va causer une donnée manquante dans le nouveau problème `tgt`. La figure 7.1 présente l'écran de saisie des symptômes de RESPIDIAG.

En deuxième lieu et après le processus RàPC, le résultat de RESPIDIAG attendu par l'utilisateur est affiché. C'est un écran comprenant le diagnostic adapté. Nous avons ajouté en fait la similarité estimée entre le nouveau problème `tgt` et le cas `srce` récupéré de la phase d'adaptation.

Dans ce qui suit, nous évaluons la performance de RESPIDIAG, selon les différentes approches proposées. Le premier chapitre présente les détails des expérimentations et de l'évaluation des premières approches proposées alors que le deuxième est consacré aux stratégies restantes.

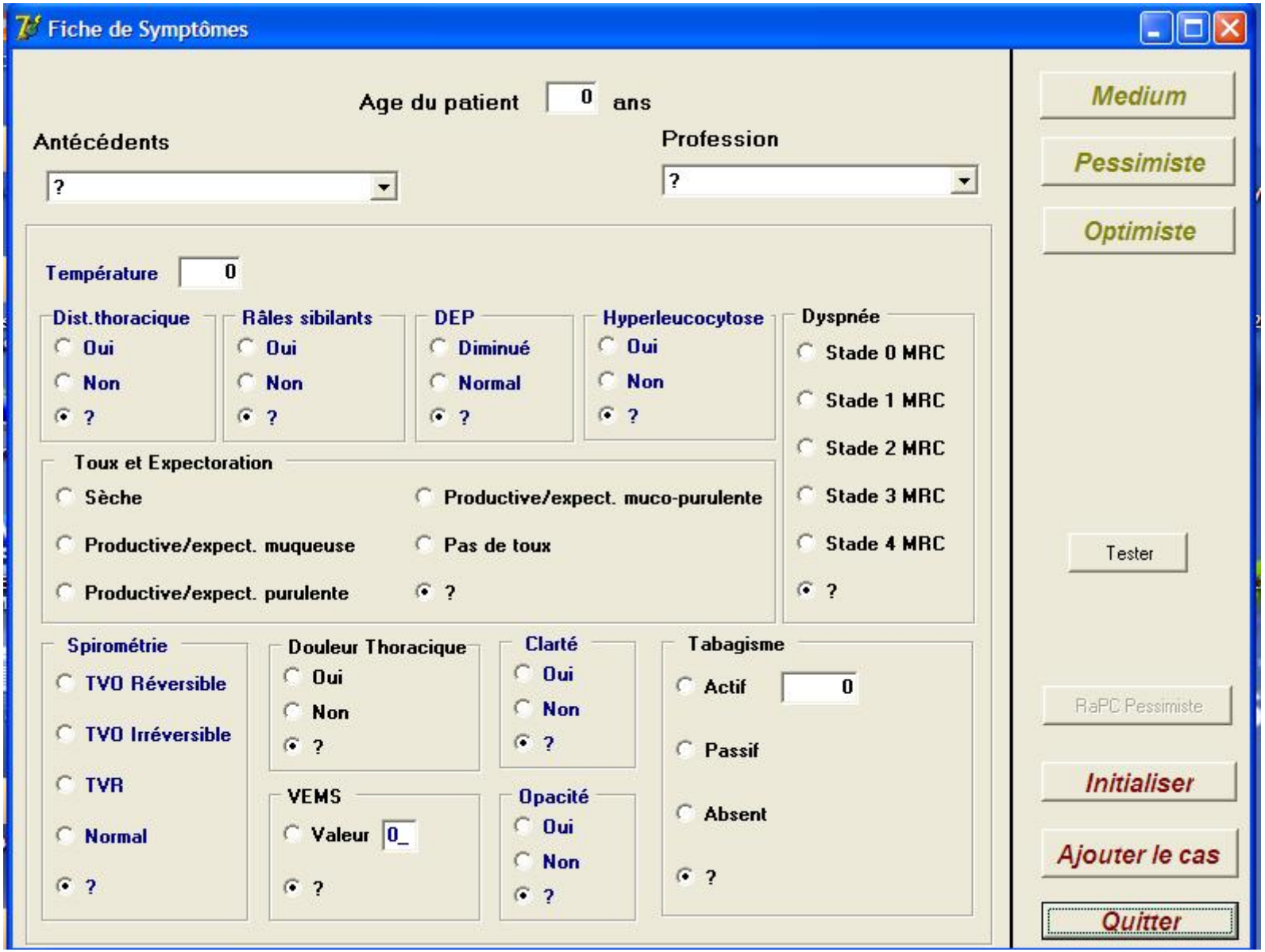

FIGURE 7.1 – Interface de saisie de RESPIDIAG.

Chapitre 8

Les premières expérimentations

Sommaire

8.1 Introduction

Dans ce chapitre nous exposons les résultats d'évaluation des quatre approches *médium*, *sélective*, *pessimiste* et *pessimiste*** dont les principes ont été présentés dans la section 5.5. L'évaluation a été réalisée sur un échantillon de test de 13 cas réels sélectionnés parmi les 21 cas dont nous avons parlé dans l'introduction de cette partie.

Les résultats de ces quatre approches récupérés *avant* la phase d'adaptation sont notés et comparés au diagnostic réel du problème `tgt`. Nous avons opté pour une première évaluation avant la phase d'adaptation, parce que nous avons voulu comparer les approches

proposées dans la phase de recherche. Par la suite, le processus de la phase d'adaptation est lancé pour avoir le résultat final de RESPIDIAG.

8.2 Résultats de l'approche *médium*

Nos premières expérimentations ont été faites sur l'approche *médium*, utilisant une similarité locale binaire -estimée selon la fonction (5.2)- pour les attributs de nature symbolique. Nous rappelons que le principe de cette approche est d'estimer moyennement la similarité locale lorsque l'une des deux valeurs concernées par le calcul est absente.

Cas	Diagnostic réel	App. *médium*	Similarité
1	1	1	0.62
2	2	4	0.81
3	2	3	0.69
4	3	4	0.74
5	1	1	0.69
6	3	3	0.75
7	5	5	0.72
8	4	4	0.79
9	4	4	0.88
10	4	4	0.86
11	5	4	0.69
12	5	4	0.68
13	4	4	0.81

TABLE 8.1 – Résultats de l'approche *médium*

La table 8.1 montre les résultats de la phase de recherche de RESPIDIAG avec cette approche. La dernière colonne représente la plus forte similarité retrouvée par RESPIDIAG

lors du processus de recherche pour le cas en question. Nous pouvons constater qu'il peut y avoir une forte similarité entre 2 cas qui n'ont pas forcément le même diagnostic, ce qui met en évidence la complication du diagnostic de la *BPCO*.

Nous avons classifié la qualité d'un résultat de la phase de recherche en fonction de la possibilité de trouver le bon (le réel) diagnostic à partir de ce résultat après la phase d'adaptation. Nous pouvons distinguer les qualités suivantes de résultats :

- *Bon* : signifie que le résultat de la phase de recherche correspond exactement au diagnostic réel du nouveau problème avant même l'adaptation.
- *Moyen* : signifie que la phase de recherche a retenu un cas source dont le diagnostic est proche de celui du nouveau problème, et que les inférences du système expert permettent d'y arriver.
- *Mauvais* : signifie que la phase d'adaptation ne permet pas de trouver le diagnostic réel du nouveau problème partant de ce résultat -diagnostic- retenu de la phase de recherche.

Le graphe 8.1 résume les résultats de cette approche en fonction de leurs qualités.

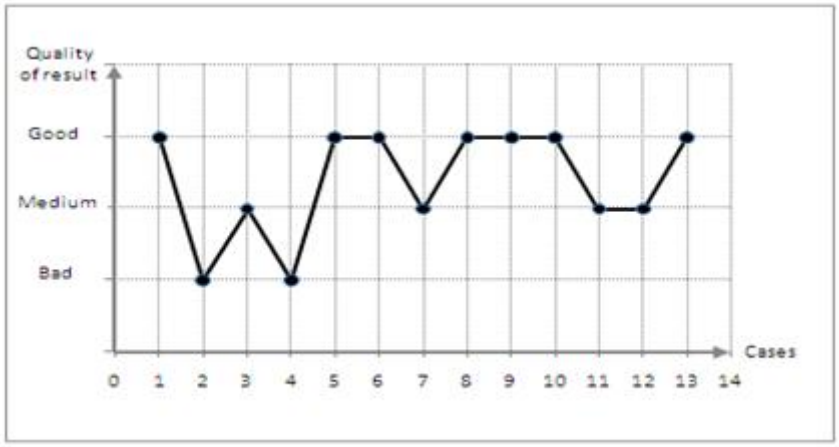

FIGURE 8.1 – Qualité des résultats de l'app. *médium*

Compte tenu de cette classification de la qualité des résultats, nous pouvons constater que nous avons obtenu avec l'approche *médium* :

– 7 *Bons* diagnostics pour les problèmes 1, 5, 6, 8, 9, 10 et 13.

– 4 *Moyens* diagnostics pour les problèmes 3, 7, 11 et 12.

– 2 *Mauvais* diagnostics pour les problèmes 2 et 4.

8.3 Résultats de l'approche *sélective*

Nous avons refait les tests sur le même échantillon précédent en utilisant l'approche *sélective* qui sélectionne uniquement les attributs renseignés dans le problème `tgt` et le cas `srce` en même temps. Les résultats sont notés dans la table 8.2. Nous pouvons constater que les similarités trouvées avec cette approche sont plus fortes que celles trouvées avec la précédente stratégie.

Cas	Diagnostic réel	*sélective*	similarité
1	1	1	0.68
2	2	4	0.85
3	2	3	0.79
4	3	3	0.80
5	1	1	0.79
6	3	3	0.87
7	5	4	0.76
8	4	4	0.86
9	4	4	0.96
10	4	4	0.92
11	5	3	0.76
12	5	4	0.70
13	4	4	0.86

TABLE 8.2 – Résultats de l'approche sélective

Cela revient principalement au fait que les attributs pour lesquels il y a des données manquantes, ont entrainé une baisse de la similarité globale car leurs similarités locales étaient estimés à 0.5 alors que pour les attributs renseignés dans les deux cas, les données sont plus rapprochées et donc offraient des similarités plus fortes. Quant à la qualité des réponses de notre système avec cette stratégie *sélective*, nous pouvons constater que nous avons obtenu :

- 8 *Bons* diagnostics pour les problèmes 1, 4, 5, 6, 8, 9, 10 et 13.
- 3 *Moyens* diagnostics pour les problèmes 3, 7 et 12.
- 2 *Mauvais* diagnostics pour les problèmes 2 et 11 .

Ces résultats sont représentés sous forme grahique dans la figure 8.2.

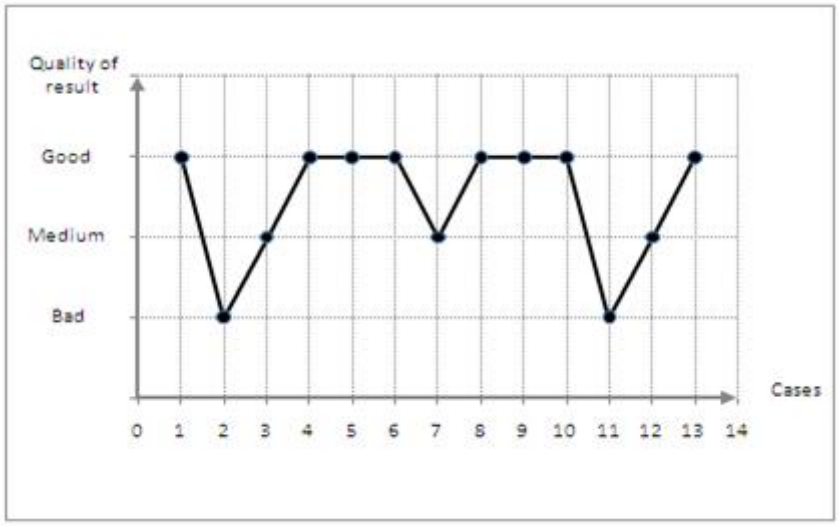

FIGURE 8.2 – Qualité des résultats de l'app. *sélective*

8.4 Résultats de l'approche *pessimiste*

L'évaluation de l'approche *pessimiste* - toujours sans l'adaptation- quant à elle a donné les résultats de la table 8.3. Nous pouvons constater que les valeurs de similarité ont légèrement diminué relativement aux résultats de l'approche *sélective*, et sont nettement

inférieures aux similarités de l'approche *médium* et cela pour l'intégralité des cas, cela revient essentiellement au principe pessimiste de l'approche.

Quant à la qualité des résultats obtenus, nous constatons une amélioration en nombre de *bons* diagnostics, et une diminution en nombre de *mauvais* diagnostics.

En effet, nous avons :

- 9 *Bons* diagnostics pour les problèmes 1, 2, 4, 5, 6, 8, 9, 10 et 13.
- 3 *Moyens* diagnostics pour les problèmes 3, 7 et 12.
- 1 *Mauvais* diagnostic pour le problème 11.

Cas	Diagnostic réel	*pessimiste*	similarité
1	1	1	0.55
2	2	2	0.67
3	2	3	0.59
4	3	3	0.61
5	1	1	0.59
6	3	3	0.67
7	5	4	0.55
8	4	4	0.65
9	4	4	0.72
10	4	4	0.72
11	5	3	0.54
12	5	4	0.52
13	4	4	0.66

TABLE 8.3 – Résultats de l'approche *pessimiste*

Ces résultats sont représentés sous forme graphique dans la figure 8.3.

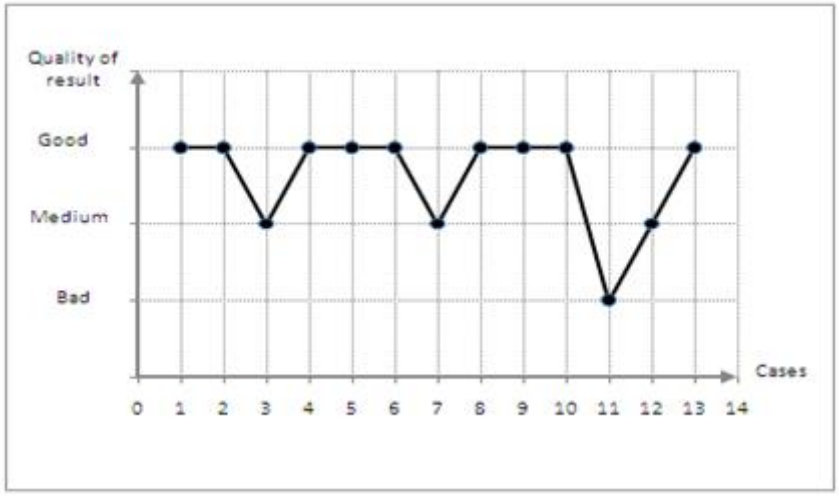

FIGURE 8.3 – Qualité des résultats de l'app. *pessimiste*

Pour récapituler les résultats de ces premières expérimentations en pourcentage, nous avons dressé la table 8.4 qui résume les qualités de résultats des trois approches *pessimiste*, *médium* et *sélective*, avec une similarité locale binaire pour les attributs symboliques, et où nous pouvons conclure facilement que la *pessimiste* présente les meilleurs résultats, suivie de la *sélective*, suivie de la *médium*.

	Bons	*Moyen*	*Mauvais*
App. *médium*	54%	31%	15%
App. *sélective*	62%	23%	15%
App. *pessimiste*	69%	23%	8%

TABLE 8.4 – Qualité des résultats des trois approches.

8.5 Résultats de l'approche *pessimiste***

En un deuxième temps d'évaluation, nous avons réappliqué l'approche *pessimiste* sur 8 nouveaux problèmes cibles, en adoptant pour les attributs symboliques, d'abords la similarité binaire (*pessimiste*) puis les similarités graduelles (*pessimiste***) proposées dans la section 5.2.3. L'objectif de cette évaluation est de tester l'apport de ces similarités graduelles sur la qualité des résultats de la phase de recherche de RESPIDIAG.

Il est à noter que le choix de l'approche *pessimiste* ici a été motivé par ses meilleurs résultats obtenus dans la section précédente comparativement aux deux autres.

La table 8.5 présente les résultats de cette évaluation.

Cas	Diagnostic réel	*pessimiste*	similarité	*pessimiste***	similarité
1	1	1	0.51	1	0.68
2	2	2	0.73	2	0.84
3	2	1	0.53	1	0.66
4	3	3	0.67	3	0.78
5	5	3	0.55	3	0.59
6	4	2	0.67	4	0.77
7	4	4	0.86	4	0.90
8	5	4	0.59	4	0.63

TABLE 8.5 – Résultats des app. *pessimiste* et *pessimiste***

Nous pouvons constater dans cette table ce qui suit :

– Pour les quatre problèmes cibles 1, 2, 4, 7, les deux approches *pessimiste* et *pessimiste*** ont donné de *bons* résultats correspondant aux diagnostics réels, mais les similarités de *pessimiste*** sont *supérieures* de 0.17, 0.11, 0.11, et 0.04 respectivement, ce qui est en faveur de la *pessimiste***.

– Pour les deux problèmes cibles 3 et 8, les *pessimiste* et *pessimiste*** donnent aussi le même diagnostic, qui est un résultat *moyen* qui peut être adapté par la phase d'adaptation pour donner le diagnostic réel. Ici également les similarités de *pessimiste*** étaient *supérieures* de 0.13, et 0.04 respectivement, ce qui est en faveur aussi de la *pessimiste***.

– Pour le problème cible 5, nos deux approches ont donné de *mauvais* résultats qui ne peuvent pas être corrigés par la phase d'adaptation. Mais l'approche *pessimiste*** donne une similarité *inférieure* à celle donnée par la *pessimiste*, un point fort pour la pessimiste**.

– Pour le problème cible Cas 6, l'approche *pessimiste* a donné un *mauvais* résultat qui ne peut pas être adapté dans la phase suivante alors que *pessimiste*** a trouvé le *bon* résultat.

Ces résultats prouvent que l'approche *pessismite*** donne de meilleurs résultats, et ceci grâce aux "fonctions heuristiques" proposées pour l'estimation des similarités entre les paires de valeurs des attributs symboliques qui ont permis de donner des résultats plus fiables et plus précis relativement aux résultats obtenus en utilisantt une fonction binaire pour les similarités locales de ces attributs.

< Nous avons appliqué ce principe de similarités graduelles uniquement dans l'approche *pessimiste*, laissant donc son application dans les deux autres approches *médium* et *sélective* pour un travail ultérieur.

8.6 Résultats de la phase d'adaptation

Comme indiqué précédemment, la phase d'adaptation du cycle RàPC de RESPIDIAG est modélisée par un système expert, dont la base de règles contient environ 30 règles permettant d'adapter le diagnostic de la phase de recherche au nouveau problème `tgt`.

Pour tester la phase d'adaptation de RESPIDIAG, nous avons repris l'ensemble des 21 cas de test et nous avons opté pour l'approche *pessimiste*** qui avait donné les meilleurs résultats jusqu'à présent.

La table 8.6 donne les résultats obtenus de cette phase d'adaptation, autrement dit, ce sont les résultats du système expert ayant reçu en entrée le diagnostic du cas retenu de la phase de recherche et comme sortie, ce même diagnostic adapté aux données du problème `tgt`.

La deuxième colonne de cette table contient les diagnostics réels des 21 cas de l'échantillon de test. Nous avons dans la troisième colonne les résultats de la phase de recherche en applicant l'approche *pessimiste*** ainsi que les similarités correspondantes aux cas retenus dans la quatrième colonne. Nous pouvons noter que nous avons obtenu :

- 13/21 bons diagnostics,
- 5/21 moyens et,
- 3/21 mauvais résultats.

La dernière colonne donne les résultats après l'adaptation, où les *bons* diagnostics restent sans changement, les *moyens* sont adaptés pour donner le diagnostic réel et les *mauvais* ne peuvent être modifiés pour la raison que nous avons déjà expliquée précédemment. Nous obtenons alors 18/21 bonnes réponses.

En conclusion nous pouvons dire que l'adaptation modélisée par un système expert donne un impact positif sur RESPIDIAG qui offre dans ces premières expérimentations jusqu'à 85% des cas le même diagnostic que celui posé par le médecin expert.

Cas	Diagnostic réel	avant	similarité	après
1	1	1	0.59	1
2	2	2	0.73	2
3	2	1	0.62	2
4	3	3	0.61	3
5	1	1	0.66	1
6	3	4	0.75	4
7	5	4	0.59	5
8	4	4	0.61	4
9	4	4	0.75	4
10	4	4	0.77	4
11	5	4	0.59	5
12	5	4	0.59	5
13	4	4	0.66	4
14	1	3	0.68	1
15	2	1	0.84	2
16	2	2	0.73	4
17	3	4	0.78	3
18	5	3	0.59	3
19	4	4	0.77	4
20	4	4	0.90	4
21	5	4	0.63	5

TABLE 8.6 – Résultats de l'approche *pessimiste*** avant et après l'adaptation.

Chapitre 9

Les deuxièmes expérimentations

Sommaire

9.1 Introduction

Dans ce chapitre nous exposons les expérimentations relatives aux cinq approches de la section 5.6. Pour la validation de nos deux approches *offline*, nous avons procédé à la duplication de la base de cas originale -contenant les données manquantes- en deux autres copies. Chacune des approches *offline* a été lancée sur une copie. Bien évidemment la duplication a été faite en structure et en contenu aussi : données présentes et données manquantes.

L'approche *statistique offline* de la section 5.6.4 a été lancée sur la première copie de la base de cas, dans l'objectif de combler les vides avec le principe de la moyenne et du vote. Nous avons obtenu en résultat un exemplaire de la base de cas originale complètement comblé.

L'approche *RàPC offline* de la section 5.6.5 quant à elle a été lancée sur la deuxième copie de la base de cas, pour combler les vides selon le principe du *RàPC*.

Nous nous retrouvons en fin de compte avec les trois bases pour lesquelles nous adopterons les notations suivantes :

– *Base A* est la base originale qui contient les vides,

– *Base B* est obtenue par application de l'approche statistique *offline* sur *Base A* et

– *Base C* est obtenue par application de l'approche *RàPC offline* sur *Base A*.

Dans ces expérimentations, deux évaluations ont été conduites. La première considère uniquement la phase de recherche, sans application de l'adaptation, elle nous a permis de comparer les différentes approches entre elles même. Et son résultat est donné dans la section 9.2. La seconde évaluation est basée sur l'utilisation des règles d'adaptation. Et son résultat est donné dans la section 9.3.

9.2 Résultats de la phase de recherche

Notre méthode d'évaluation dans cette partie, consiste à exécuter chacune des trois approches *online* : *pessimiste**, *médium** et *optimiste* des sections 5.6.1, 5.6.3 et 5.6.2 respectivement, sur chacune des bases : *Base A*, *Base B* et *Base C*. Sachant que les bases B et C sont totalement comblées, l'utilisation des approches *online* reste indispensable pour prendre en charge le manque des données au niveau des problèmes $cibles$.

Il est à noter que chaque exécution sous entend le test des 21 problèmes $cibles$ de notre échantillon de test. Donc en fin de compte, $3(approches) \times 3(bases) \times 21(cas) = 189$ diagnostics sont générés à la base de :

$17(sympt.) \times 21(cas) \times 3(approches) \times 3(bases) = 3213$ symptômes saisis.

Ensuite, ces diagnostics sont comparés avec les diagnostics réels des 21 patients de test.

Il est à noter également que ce travail ne signifie pas que RESPIDIAG possède trois bases de cas en même temps. En réalité et au niveau de l'implémentation, nous changeons à chaque fois le lien de RESPIDIAG avec sa base de cas. Ainsi, nous lions au début le système à la base de cas A (par exemple), nous activons l'appel d'une approche online bien spécifique, puis nous exécutons les tests des 21 problèmes $cibles$ et nous notons les résultats. A la fin de cette étape, nous désactivons l'appel de cette approche, pour activer l'appel d'une deuxième approche online tout en gardant RESPIDIAG liée à la même base A. Nous refons les tests, puis nous procédons de la même manière pour la troisième approche. A la fin de ce travail, il est temps de modifier le lien de la base, qui se fait cette fois ci vers la deuxième base B (par exemple), et le processus est répété jusqu'à obtention de tous les résultats.

Les tables 9.1, 9.2 et 9.3 donnent les résultats de l'application de nos trois approches sur les *bases A, B et C* respectivement. Dans ces tables, nous retrouvons le diagnostic réel du problème $cible$, les trois diagnostics générés par les différentes stratégies pour le même problème accompagné à chaque fois de la similarité (la plus forte trouvée) entre ce problème et le cas $source$ retenu.

Sur la base des trois tables 9.1, 9.2 et 9.3, nous avons dressé les tables 9.4 et 9.5 qui donnent une synthèse des nombres de réponses correctes, par approche et par base.

Dans ces expérimentations et sur les trois bases de cas, nous constatons que la moyenne des résultats de l'approche *pessimiste** est strictement meilleure que la moyenne des résultats de l'approche *médium** qui est, à son tour strictement meilleure que l'approche *optimiste*.

Nous pouvons constater également que la meilleure approche est d'utiliser la base de cas Base C complétée grâce à l'approche *RàPC offline* et de prendre en compte le

N	Diag. réel	*optimiste*	Sim	*pessimiste**	Sim	*médium**	Sim
1	1	3	0.91	1	0.53	1	0.67
2	2	1	0.92	3	0.72	3	0.82
3	2	1	0.95	1	0.58	1	0.75
4	3	3	0.96	3	0.58	3	0.76
5	1	1	0.89	1	0.76	1	0.79
6	3	3	0.94	3	0.77	3	0.81
7	5	5	0.95	5	0.76	5	0.82
8	4	5	1	3	0.62	3	0.79
9	4	2	0.99	4	0.73	4	0.83
10	4	3	0.94	4	0.73	4	0.82
11	5	5	1	5	0.71	5	0.81
12	5	5	0.95	5	0.70	5	0.77
13	4	3	0.91	4	0.72	4	0.80
14	1	1	0.85	1	0.59	1	0.70
15	2	2	0.92	4	0.81	4	0.84
16	2	3	0.92	1	0.68	3	0.76
17	3	3	0.95	3	0.71	3	0.77
18	5	5	0.94	5	0.77	5	0.83
19	4	5	0.98	4	0.81	3	0.83
20	4	2	0.97	4	0.89	4	0.91
21	5	5	1	5	0.80	5	0.86

TABLE 9.1 – Résultats des approches *online* sur la base A

N	Diag. réel	*optimiste*	Sim	*pessimiste**	Sim	*médium**	Sim
1	1	1	0.80	1	0.62	1	0.71
2	2	3	0.90	3	0.74	3	0.83
3	2	2	0.91	2	0.66	2	0.79
4	3	3	0.90	3	0.69	3	0.80
5	1	1	0.86	1	0.82	1	0.84
6	3	3	0.85	3	0.85	3	0.85
7	5	1	0.93	5	0.82	5	0.86
8	4	3	0.96	4	0.63	3	0.78
9	4	4	0.97	4	0.76	4	0.86
10	4	4	0.93	4	0.73	4	0.82
11	5	3	0.92	5	0.79	5	0.86
12	5	5	0.90	5	0.78	5	0.83
13	4	4	0.88	4	0.72	4	0.80
14	1	1	0.83	1	0.73	1	0.78
15	2	3	0.90	3	0.85	3	0.88
16	2	3	0.86	3	0.76	3	0.81
17	3	3	0.81	3	0.79	3	0.80
18	5	3	0.90	5	0.82	5	0.86
19	4	3	0.88	3	0.83	3	0.85
20	4	4	0.94	4	0.92	4	0.93
21	5	5	0.95	5	0.88	5	0.91

TABLE 9.2 – Résultats des approches *online* sur la base B

N	Diag. réel	*optimiste*	Sim	*pessimiste**	Sim	*médium**	Sim
1	1	1	0.80	1	0.62	1	0.71
2	2	3	0.90	3	0.74	3	0.83
3	2	2	0.91	2	0.66	2	0.79
4	3	3	0.88	3	0.67	3	0.78
5	1	1	0.86	1	0.82	1	0.84
6	3	3	0.85	3	0.85	3	0.85
7	5	1	0.93	5	0.84	5	0.88
8	4	3	0.96	4	0.63	3	0.78
9	4	4	0.97	4	0.76	4	0.86
10	4	4	0.93	4	0.73	4	0.82
11	5	5	0.95	5	0.82	5	0.88
12	5	5	0.93	5	0.80	5	0.85
13	4	3	0.88	4	0.72	4	0.80
14	1	1	0.83	1	0.73	1	0.78
15	2	3	0.90	3	0.85	3	0.88
16	2	3	0.85	3	0.74	3	0.79
17	3	3	0.80	3	0.77	3	0.78
18	5	3	0.90	5	0.82	5	0.86
19	4	4	0.90	4	0.83	3	0.86
20	4	4	0.94	4	0.92	4	0.93
21	5	5	0.97	5	0.91	5	0.94

TABLE 9.3 – Résultats des approches *online* sur la base C

	Base A	Base B	Base C
Approche *optimiste*	11	13	14
Approche *pessimiste**	16	17	18
Approche *médium**	15	16	16

TABLE 9.4 – Nombres des réponses correctes

problème du manque de données dans le problème cible par l'utilisation de l'approche *pessimiste* online* qui atteint les 18/21 réponses correctes sur cette base.

Finalement, ces résultats sont utiles aussi pour comparer les trois stratégies pour le comblement dans les bases de cas. La pire stratégie dans ces expérimentations consiste à ne pas combler les vides, laissant ainsi la gestion des données manquantes entièrement aux approches *online*.

Donc il est claire que le comblement des bases B et C par nos approches *offline* a amélioré la qualité des résultats des approches *online*, et que la base C permet d'avoir le taux le plus élevé de réponses correctes qui atteint 76.19 %.

	Nombre	Pourcentage
Base A	42	66.66%
Base B	46	73.60%
Base C	48	76.19%

TABLE 9.5 – Nombres de réponses correctes par base

Les données de cette table sont représentées graphiquement dans la figure 9.1.

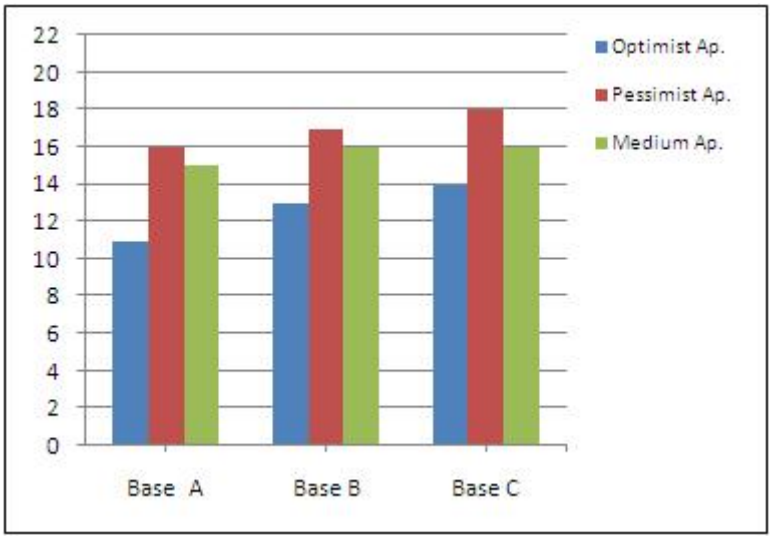

FIGURE 9.1 – Nombres de réponses correctes par base

9.3 Résultats après la phase d'adaptation

Dans cette section, nous présentons les résultats des approches *online* de la section précédente après le processus de l'adaptation. Nous rappelons que ce processus a été exécuté suite au processus de recherche, et nous nous retrouvons avec 189 diagnostics adaptés. Les résultats sont présentés ci-dessous dans les tables 9.6, 9.7 et 9.8 et correspondent respectivement aux résultats sur les bases A, B et C. Ces mêmes résultats sont synthétisés dans la table 9.9 qui montre que les conclusions tirées précédemment à propos de nos approches proposées avec seulement la phase de recherche restent valables aussi après l'adaptation. En particulier, la meilleure approche est d'utiliser la base de cas C complétée grâce à l'approche *RàPC offline* et d'utiliser l'approche *pessimiste** pour la prise en charge du manque de l'information dans le problème $cible$. En effet, nous nous retrouvons avec $21/21$ réponses correctes du système avec cette approche, suivie toujours de la même approche *pessimiste** mais sur la base B ($20/21$ bonnes réponses), suivie de ses 19 bonnes réponses sur la base A, et de la *médium** sur les deux bases B et C. L'approche *optimiste* quant à elle donne les plus faibles résultats avec 17, 16 et 17 bonnes réponses obtenues sur les bases A, B et C respectivement.

N	Diag. réel	*optimiste*	après	*pessimiste**	après	*médium**	après
1	1	3	1	1	1	1	1
2	2	1	2	3	2	3	2
3	2	1	2	1	2	1	2
4	3	3	3	3	3	3	3
5	1	1	1	1	1	1	1
6	3	3	3	3	3	3	3
7	5	5	5	5	5	5	5
8	4	5	4	3	3	3	3
9	4	2	2	4	4	4	4
10	4	3	2	4	4	4	4
11	5	5	5	5	5	5	5
12	5	5	5	5	5	5	5
13	4	3	2	4	4	4	4
14	1	1	1	1	1	1	1
15	2	2	2	4	4	4	4
16	2	3	2	1	2	3	2
17	3	3	3	3	3	3	3
18	5	5	5	5	5	5	5
19	4	5	4	4	4	3	3
20	4	2	3	4	4	4	4
21	5	5	5	5	5	5	5

TABLE 9.6 – Résultats des approches *online* sur la base A après l'adaptation

N	Diag. réel	*optimiste*	après	*pessimiste**	après	*médium**	après
1	1	1	1	1	1	1	1
2	2	3	2	3	2	3	2
3	2	2	2	2	2	2	2
4	3	3	3	3	3	3	3
5	1	1	1	1	1	1	1
6	3	3	3	3	3	3	3
7	5	1	1	5	5	5	5
8	4	3	3	4	4	3	3
9	4	4	4	4	4	4	4
10	4	4	4	4	4	4	4
11	5	3	3	5	5	5	5
12	5	5	5	5	5	5	5
13	4	4	4	4	4	4	4
14	1	1	1	1	1	1	1
15	2	3	2	3	2	3	2
16	2	3	2	3	2	3	2
17	3	3	3	3	3	3	3
18	5	3	3	5	5	5	5
19	4	3	3	3	3	3	3
20	4	4	4	4	4	4	4
21	5	5	5	5	5	5	5

TABLE 9.7 – Résultats des approches *online* sur la base B après l'adaptation

N	Diag. réel	*optimiste*	après	*pessimiste**	après	*médium**	après
1	1	1	1	1	1	1	1
2	2	3	2	3	2	3	2
3	2	2	2	2	2	2	2
4	3	3	3	3	3	3	3
5	1	1	1	1	1	1	1
6	3	3	3	3	3	3	3
7	5	1	1	5	5	5	5
8	4	3	3	4	4	3	3
9	4	4	4	4	4	4	4
10	4	4	4	4	4	4	4
11	5	5	5	5	5	5	5
12	5	5	5	5	5	5	5
13	4	3	3	4	4	4	4
14	1	1	1	1	1	1	1
15	2	3	2	3	2	3	2
16	2	3	2	3	2	3	2
17	3	3	3	3	3	3	3
18	5	3	3	5	5	5	5
19	4	4	4	4	4	3	3
20	4	4	4	4	4	4	4
21	5	5	5	5	5	5	5

TABLE 9.8 – Résultats des approches *online* sur la base C après l'adaptation

	Base A	Base B	Base C
App. *optimiste*	17	16	17
App.*pessimiste**	19	20	21
App. *médium**	18	19	19

TABLE 9.9 – Nombres de réponses correctes après le processus d'adaptation

La figure 9.2 présente ces résultats sous forme graphique, et sa comparaison avec la figure 9.1 fait sortir aussi que le processus d'adaptation modélisé par un système expert a bien amélioré la performance de RESPIDIAG : pour chacune des 3×3 configurations (3 bases de cas et 3 approches *online*), la moyenne du résultat après l'adaptation est meilleure que la moyenne des résultats avant l'adaptation (strictement meilleure pour 8 parmi eux). Cela montre le rôle très positif de notre modélisation de l'adaptation par un système expert.

FIGURE 9.2 – Résulats des différentes approches par base et après l'adaptation

Discussion et travaux en relation

Les résultats des évaluations présentés dans les deux derniers chapitres, montrent bien que les apports de cette thèse sur le plan traitement du missing data, ou bien sur le plan adaptation, sont très bénéfiques pour la performance de notre système RESPIDIAG, qui vient s'ajouter à une longue liste de systèmes médicaux basés sur le principe du *RàPC*. Mais en réalité, RESPIDIAG s'ajoute aussi à une autre catégorie de systèmes basés sur d'autres techniques de l'intelligence artificielle mais qui traitent comme RESPIDIAG le problème du missing data.

Nous avons mentionné déjà quelques uns de ces systèmes dans la section 3.4. A titre d'exemples, nous prenons le travail de [Lin08] où les auteurs proposent un réseau bayésien pour la prédiction des problèmes médicaux. Leur première méthode consiste simplement à faire l'apprentissage du réseau sans aucun pré-traitement des valeurs manquantes. Ceci ressemble bien à notre première stratégie dans RESPIDIAG, qui consiste à ne pas combler les vides et à laisser entièrement la gestion des données manquantes aux approches *online*. Leur seconde méthode consiste à calculer la valeur manquante par la moyenne globale de toutes les valeurs disponibles dans l'ensemble de données de leur réseau. Et ceci peut ressembler à la première approche *offline* dans RESPIDIAG qui utilise la moyenne des valeurs présentes mais enrichie encore avec le principe du vote pour les valeurs des attributs de type logique ou symbolique.

Dans [Pesonen98], le problème du missing data considéré dans un système de réseau de neurones -diagnostic des douleurs abdominales aigues - est à propos d'un seul attribut

uniquement : le nombre des leucocytes. La charge est nettement plus lourde dans RESPIDIAG où nous traitons le problème du missing data dans 17 attributs à la fois et qui sont en plus de *différentes natures*, et où les données manquantes apparaissent de manière aléatoire dans chaque cas.

Les différentes méthodes exposées dans [Pesonen98] pour le comblement des vides montrent des écarts importants dans les valeurs substituées, à l'exception des méthodes du plus proche voisin et le réseau de neuronnes qui donnent des valeurs plus ou moins proches. Dans RESPIDIAG, nos deux approches *offline* convergent dans les valeurs substituées malgré leurs principes différents et ce pour la majoroité des cas. Cela prouve bien une fiabilité quant aux valeurs proposées par nos méthodes.

Dans [Ocompo11], un travail récent portant sur une comparaison des inférences bayésiennes et l'approche RàPC appliqué au diagnostic de méningites peut être trouvé. La comparaison montre que le système RàPC donne des résultats plus précis que le système bayésien. Le même travail compare les résultats du système RàPC avant et après l'adaptation, et conclue que la phase d'adaptation donne plus de flexibilité et plus de robustesse au système. Dans RESPIDIAG également, l'adaptation a bien amélioré les résultats de la phase de recherche, en atteignant le 100% de réponses correctes dans le test des 21 cas réels, et avec l'approche *online pessimiste** lancée sur la base C qui a été comblée par l'approche *offline RàPC*.

Conclusion et Perspectives

Cette thèse présente RESPIDIAG, un système d'aide à la décision médicale basé sur le principe du raisonnement à partir de cas, et dédié au diagnostic de la Broncho Pneumopathie Chronique Obstructive. Cette dernière est une maladie très dangereuse causée par le tabac, qui touche généralement les fumeurs ou ex-fumeurs agés de plus de 50 ans. Le travail considère la *BPCO* dans son état de base avec ses quatre stades de sévérité, en plus de deux exacerbations possibles de la *BPCO* qui peuvent être d'origine infectieuse ou d'origine pneumo-thorax. Ce qui donne les six diagnostics de notre champ d'étude. Un cas de cette application décrit un patient grâce à 17 attributs numériques, logiques et symboliques, et un diagnostic choisi parmi les six.

Le premier objectif de cette thèse est donc la réalisation de ce système avec l'implémentation de toutes les phases d'un cycle de RàPC. La première de ces phases est basée sur une mesure de similarité $\mathcal{S}$ définie comme étant la moyenne pondérée des mesures de similarités locales $\mathcal{S}_a$ associées à chaque attribut a. La difficulté principale pour cette application est liée aux données manquantes dans la base de cas avec environ 21%, et dans le problème cible avec environ 11% dans l'échantillon de test. Ces données manquantes empêchent le calcul des similarités locales, et par conséquent la similarité globale ne peut avoir lieu.

Le deuxième objectif de cette thèse consiste alors à une prise en charge de l'incertain dans RESPIDIAG. Nous entendons par cette prise en charge la gestion de deux situations différentes : raisonner malgré des données manquantes ou bien raisonner pour combler

les données manquantes. Dans RESPIDIAG, ces données absentes peuvent être observées dans le nouveau problème à traiter ou bien dans les anciens cas sources de la base de cas.

Dans l'ordre de dresser une issue, quatre stratégies *online* et deux stratégies *offline* sont définies. Les approches *online* visent à attribuer, lors de l'exécution du processus de RESPIDIAG, particulièrement pendant la phase de recherche, une valeur à la similarité locale de l'attribut pour lequel une au moins des valeurs dans le cas $source$ ou dans le problème $cible$ est inconnue. Les stratégies *offline* quant à elles visent un autre objectif : combler les vides laissés par les données manquantes dans la base de cas, elles ne concernent pas l'absence des données au niveau des problèmes $cibles$, qu'elles confient totalement aux approches *online*.

Au début, trois approches *online* ont été proposées à savoir, la *pessimiste*, la *médium* et la *sélective*. Des évaluations ont été réalisées selon lesquelles la *pessimiste* avait donné les meilleurs résultats. Raison pour laquelle une variante améliorée de cette approche a été définie par la suite, qui prend alors pour nom *pessimiste***. Cette dernière a été évaluée également, et elle a prouvé un impact positif sur les résultats.

Dans un deuxième temps, deux stratégies *offline* ont été proposées, à savoir, l'approche *statistique* et l'approche *RàPC offline*. Elles ont comblé le vide de la base de cas originale selon deux principes différents. Sur un autre plan, des variantes améliorées ont été proposées pour les approches *pessimiste*** et *médium*, pour obtenir finalement la *pessimiste** et la *médium**. Une autre approche appelée *optimiste* a été définie à ce niveau aussi.

La dernière évaluation relative à la phase de recherche a été d'appliquer ces trois dernières approches sur les trois bases de cas obtenues. Elle a montré que l'approche *pessimiste** a généré le meilleur résultat sur l'échantillon de test, lorsque la base de cas a été complétée par le principe de l'approche *RàPC offline*.

Pour la phase d'adaptation, nous avons intégré un autre type de raisonnement pendant

le processus *RàPC* de RESPIDIAG. Il s'agit bien d'un système expert à base de règles de production qui se charge d'adapter le diagnostic récupéré du cas $source$ retenu dans la phase de recherche aux données du nouveau patient. L'idée de cette intégration a été motivée principalement par la performance des systèmes experts dans la modélisation du raisonnement que l'expert peut expliquer. Et en effet, les résultats de cette phase ainsi modélisée ont montré une nette amélioration de leurs qualités.

Bien que RESPIDIAG donne déjà des résultats satisfaisants, il y a place à l'amélioration. Donc des perspectives sont planifiées. Une partie d'elles est déjà mentionnée dans le corps de la thèse, d'autres sont présentées ci-dessous.

La première visera à prendre en compte la connaissance du domaine pour les stratégies *online*. La connaissance du domaine considérée est donnée par contraintes entre les attributs, affirmant que certaines combinaisons d'attributs ne sont pas licites. Par exemple, dans un autre domaine d'application, on peut dire que la hauteur d'une personne est handicapée par son âge et son sexe. Par exemple, la hauteur d'un garçon de 8 ans est nécessairement inférieure à, disons, 150 centimètres.

Par conséquent, au lieu d'utiliser la plage de données d'un attribut pour estimer sa valeur manquante en utilisant des stratégies *optimiste, pessimiste* ou *médium*, un sous-ensemble de sa plage entraîné par les contraintes de la connaissance du domaine peut être utilisé, et qui fournira des valeurs estimées qui peuvent être plus précises et donc, devrait permettre d'améliorer le système. Ce travail futur soulève également des questions d'acquisition de connaissances du domaine de chez les médecins experts.

Dans la section 5.6.5, nous avons noté que les poids w_b^a peuvent être différents des poids w_b et des poids w_b^c ($c \neq a$), respectivement utilisés pour le *"processus RàPC pour l'estimation de* a*"*, le *"processus RàPC pour le diagnostic de la* $BPCO$*"*, et le *"processus RàPC pour l'estimation de* c*"*. Cependant, pour nos expérimentations, nous avons choisi les mêmes valeurs : pour chaque $a, b \in$ `Attributs` tel que $a \neq b$, $w_b = w_b^a$. En effet,

l'estimation des poids est coûteuse en terme du temps expert.

Un futur travail consistera à trouver un chemin pour mieux estimer les poids w_b^a sans trop additionner du temps de l'expert. Une autre manière de le faire pourrait être d'appliquer une approche d'apprentissage par renforcement, à l'aide du score de l'évaluation sur l'échantillon de test : une modification de l'ensemble des poids serait évaluée par la variation de cette partition.

Toujours pour la phase de recherche de RESPIDIAG, nous avons utilisé la méthode des k-plus proches voisins avec $k = 1$. Une autre perspective pouvant bien être intéressante est de retester les mêmes stratégies du traitement de données manquantes en augmentant la valeur de k, à 1, 2 ... et de comparer les résultats. Une certaine valeur de k, autre que 1, pourrait donner peut être des résultats plus intéressants à RESPIDIAG.

Dans nos premières expérimentations, la phase d'adaptation a été lancée uniquement sur l'approche *pessimiste*** ayant donné les meilleurs résultats. Nous avons suivi en fait, une théorie qui dit que plus la solution récupérée de la phase de recherche, est proche du cas cible, plus l'adaptation fournit un résultat meilleur. Une autre théorie, dit que le meilleur cas récupéré de la phase de recherche, n'est pas forcément le meilleur à adapter pour avoir un résultat adequat au problème cible : nous nous proposons alors d'essayer dans le futur l'adaptation par le système expert sur toutes les autres approches de nos premières expérimentations, à savoir la *pessimiste*, la *médium* et la *sélective*.

Une autre perspective est focalisée sur la base de cas de RESPIDIAG. Le fait d'augmenter son volume pourra bien donner des résultats plus significatifs. Néanmoins, cette tâche est très couteuse en terme du temps des experts, surtout que les dossiers des patients au niveau du service de pneumologie sont encore sous forme de papiers. Il est à noter que mettre l'archive du service sous forme informatisée est en lui-même un projet à part, vu le temps expert recquis pour cette tâche.

Annexe A

R1	**if** $\texttt{sol}(\texttt{srce}) = \texttt{Stade\_I}$ **and** $\texttt{tgt.VEMS} \leq 30\%$ **then** $\texttt{sol(tgt)} = \texttt{Stade\_IV}$
R2	**if** $\texttt{sol}(\texttt{srce}) = \texttt{Stade\_II}$ **and** $\texttt{tgt.VEMS} \leq 50\%$ **and** $\texttt{tgt.VEMS} \succcurlyeq 30$ **then** $\texttt{sol(tgt)} = \texttt{Stade\_III}$
R	**if** $\texttt{sol}(\texttt{srce}) = \texttt{ExPnThor}$ **and** $\texttt{tgt.hyperLeucocytose} = yes$ **then** $\texttt{sol(tgt)} = \texttt{ExInfec}$
R	**if** $\texttt{sol}(\texttt{srce}) = \texttt{ExInfec}$ **and** $\texttt{tgt.douleurThoracique} = yes$ **then** $\texttt{sol(tgt)} = \texttt{ExPnThor}$

TABLE A.1 – Les règles d'adaptations

Bibliographie

[Aamodt94] A. Aamodt, *Case-based reasoning : Foundational issues, methodological variations, and system approaches* proceeding of AICOM'94, pp. 39–58.

[Anane04] T. Anane, M. Baghriche, R. BenAli, R. Boukari, Y. Berrabah, P. Chaulet, A. Fissah, D. Larbaoui, S. Lellou, S. Nafti and N. Zidouni, *Guide pour la prise en charge des maladies respiratoires de l'enfant et de l'adulte dans les unités sanitaires de base en Algérie*, 2004, Institut National de Santé Publique.

[Balaa03] Z. E. Balaa, A. Strauss, P. Uziel, K. Maximini and R. Traphöner, *FM-Ultranet : A decision support system using case-based reasoning applied to ultrasonography*, in L. McGinty, (Ed.) : Workshop Proceedings of the International Conference on Case-Based Reasoning ICCBR-2003, pp 37–44.

[Bichindaritz10] I. Bichindaritz and C. Marling, *Case-Based Reasoning in the Health Sciences : Foundations and Research Directions*, Computational Intelligence in Healthcare, vol 4, SCI 309, 2010, pp 127–157.

[Bradburn93] C. Bradburn and J. Zeleznikow, *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, Vol.837 pp. 365–378, 1993.

[Chae96] Y.M. Chae, *chapter Expert Systems in medicine*, the Handbook of Applied Expert Systems, pp 767-790, 1996.

[Fabbri11] L. M. Fabbri, L. Fabrizio, B. Beghe and K. F. Rabe, *The Multiple Components of COPD. COPD : A Guide to Diagnosis and Clinical Management, Respiratory Medicine* DOI 10.1007/978-1-59745-357-8-1, LLC (2011).

[Fuchs00] B. Fuchs et A. Mille *Une modélisation au niveau connaissance du raisonnement à partir de cas*, IC'2000, Journées francophones d'ingénierie des connaissances, Toulouse , FRANCE (10/05/2000), 2000, pp. 3-11.

[Gierl98] L. Gierl, M. Bull and R. Schmidt, *CBR in Medicine*, Proceeding of Case-Based Reasoning Technology, From Foundations to Applications Pages 273-298.

[Guessoum06] S. Guessoum et M.T. Laskri, *Couplage de SBC/CBR pour la prise en charge de la santé respiratoire*, MCSEAI'06, Maghrebian Conference on Software Engineering and Artificial Intelligence, Agadir, Maroc, Décembre 2006.

[Guessoum07] S. Guessoum, M.T. Laskri et R. Benali, *Respidiag : un système multi-agents pour la prise en charge de la santé respiratoire*, BIIA, Bulletin d'Informatique Approfondie et Applications de l'Université de Provence France, pages 3- 8, num 77 Juin 2007, ISSN 0291-5413.

[Guessoum11] S. Guessoum, N. Dendani et H. Djellali, *Vers une Approche de Recherche utilisant le Raisonnement à partir de cas appliqué au diagnostic de la Broncho Pneumopathie Chronique Obstructive*, Proceeding du 19ème Atelier du Raisonnement à Partir de Cas, Plate Forme AFIA, Mai 2011, Chambéry, France.

[Guessoum12-a] S. Guessoum, M.T. Laskri, H. Djellali and M.T. Khadir, *Combining Case and Rule Based Reasoning for the Diagnosis and Therapy of Chronic Obstructive Pulmonary Disease*, international journal of hybrid information technology, vol 5, issue 3, july 2012.

[Guessoum12-b] S. Guessoum and M.T. Laskri, *Case-Based Reasoning System for Medical Diagnosis*, Proceeding of BIOMEIC'12, Biomedical Engineering International Conference, Octobre 2012, Telemcen Algérie.

[Guessoum12-c] S. Guessoum and M.T. Laskri, *Case-Based Support for the Diagnosis of Chronic Obstructive Pulmonary Disease*, Proceeding of ICCS'12, International Conference on Complex Systems, IEEE, Novembre 2012, Agadir, Maroc.

[Guessoum13] S. Guessoum, M.T. Laskri and J. Lieber, *RESPIDIAG : un système de raisonnement à partir de cas pour le diagnostic de la broncho-pneumopathie chronique obstructive*, Proceeding du 21ème Atelier du Raisonnement à Partir de Cas dans IC'13, Conférence Internationale de l'Ingénierie de Connaissances, Juillet 2013, Lille, France.

[Guessoum14] S. Guessoum, M.T. Laskri and J. Lieber *RespiDiag : a case-based reasoning system for the diagnosis of chronic obstructive pulmonary disease*, journal of Expert systems with applications, vol 41, issue 2, pp : 267–273.

[Haddad97] M. Haddad, K.P. Adlassnig and G. Porenta, *Feasability analysis of a case-based reasoning system for automated detection of coronary heart disease from myocardial scintigrams*, journal of Artificial Intelligence in Medicine 9 (1997) 61–78.

[Jurisica98] I. Jurisica, J.Mylopoulos, J. Glasgow, H. Shapiro and RF. Casper, *Case-based reasoning in IVF : Prediction and knowledge mining*, journal of Artificial Intelligence in Medicine, 1998, vol 12, issue 1, pp 1–24.

[Kahn94] Kahn CE Jr *Clinical trial and evaluation of a prototype case-based system for planning medical imaging work-up strategies*, In Case-Based Reasoning : Papers from the Workshop, Menlo Park, CA, AAAI Press, 1994, p 138.

[Koton88] P. Koton, *Reasoning about evidence in causal explanations*, In : J. Kolodner, (Ed.) : Proceedings of the Case-Based Reasoning Workshop, Clearwater Beach, Florida 1988, pp 260–270.

[Lieber00] J. Lieber and B. Bresson, *Case-based reasoning for breast cancer treatment decision helping*, in E. Blanzieri and L. Portinale, (Eds.) : Advances in case-based reasoning. Proceedings of the European Workshop on Case-Based Reasoning, EWCBR. Lecture Notes in Artificial Intelligence, Vol. 1898, pp. 173–185 (2000).

[Lin08] J.H. Lin and P. J. Haug, *Exploiting missing clinical data in Bayesian network modeling for predicting medical problems*, Journal of Biomedical Informatics, vol 41, 2008, pp 1-14.

[Mantaras96] R.L.D. Mantaras and E. Plaza, *Case Based Reasoning : An Overview*, Artificial Intelligence Research Institute. CSIC-Spanish National Reaserch Council.

[Mille99] A. Mille, *Etat de l'art du raisonnement à partir de cas*, Plate forme AFIA, 1999, Palaiseau, France.

[Nuovo11] A.G.Di Nuovo, *Missing data analysis with fuzzy C-Means : A study of its application in a psychological scenario*, journal of Expert Systems with Applications, vol 38, 2011, pp 6793-6797.

[Ocompo11] E. Ocampo, M. Maceiras, S. Herrera, C. Maurente, D. Rodriguez and M. Sicilia, *Comparing Bayesian inference and case-based reasoning as support techniques in the diagnosis of Acute Bacterial Meningitis*, journal of Expert Systems with Applications, vol 38, 2011, pp 10343-10354.

[Opiyo10] E.T.O. Opiyo. *Case-based reasoning for expertise relocation in support of rural health workers in developing countries*, in A. Aamodt and M. Veloso (Eds.) : Case-based reasoning research and development. Proceedings of the International Conference on Case-Based Reasoning, ICCBR-95. Lecture Notes in Artificial Intelligence, Vol. 1010, 1995. 77–87.

[Pesonen98] E. Pesonen, M. Eskelinen and M. Juhola, *Treatment of missing data values in a neural network based decision support system for acute abdominal pain*, journal of Artificial Intelligence in Medicine, vol 13, 1998, pp 139-146.

[Schmidt01] R. Schmidt and L. Gierl, *A prognostic model for temporal courses that combines temporal abstraction and case-based reasoning*, Journal of Medical Informatics 74(2-4) : 307-315 (2005).

[Schmidt07] R. Schmidt, *Case-Based Reasoning in Medicine Especially an Obituary on Lothar Gierl*, Studies in Computational Intelligence (SCI) 48, (2007) pp. 63–87.

[Schmidt09] R. Schmidt and O. Vorobieva, *Combining Statistics and Case-Based Reasoning for Medical Research*, C.L. Mumford and L.C. Jain (Eds.) : Computational Intelligence, ISRL 1, pp. 673–696. (2009).

[Sweb] `www.capitalsouffle.fr/COPD.htm`, last consultation, September 2012,

[Thomesse04] J.P. Thomesse, F. Charpillet, L. Véga, P. Durand and J. Chanliau *Diatélic : un système intelligent de télémédecine appliqué à la dialyse : aspects médicaux, informatiques, économiques*, Institut National Polytechnique de Lorraine, Nancy, France et ALTIR, Association Lorraine pour le Traitement de l'Insuffisance Rénale, 2004.

Résumé

Dans cette thèse, un système d'aide à la décision médicale est présenté. Il est dédié au diagnostic d'une trés grave maladie respiratoire causée par le tabac et nommée la Broncho-Pneumopathie Chronique Obstructive (ou BPCO). Le système est basé sur les principes de raisonnement à partir de cas et réunie l'expérience de plusieurs médecins experts du service de pneumologie de l'hôpital Dorban (Annaba, Algérie).

Une question essentielle à propos de la base de cas est que certaines valeurs d'attributs sont absentes dans la plupart des cas. Dans ce contexte de données manquantes, quatre approches avec certaines variantes ont été proposées, implémentées et évaluées. Elles visent l'estimation de la similarité locale entre deux valeurs lorsque l'une d'elles au moins est manquante, et s'exécutent pendant le processus *online* du système. Deux autres approches ont été proposées également pour combler les vides laissés par les valeurs manquantes par des valeurs plausibles, en utilisant une méthode statistique et une autre basée sur le principe du raisonnement à partir de cas lui- même. Une autre contribution de cette thèse est la modélisation de la phase d'adaptation par un système expert à base de règles de production utilisant le chainage avant. Cette intégration est motivée par la performance des systèmes experts à modéliser les raisonnements que l'expert humain arrive à expliquer. L'objectif principal étant de rendre les résultats du système les plus fiables possibles.

Mots-clés: Raisonnement à partir de cas, raisonnement à base de règles, diagnoctic de la BPCO, aide à la décision médicale, les données manquantes

Abstract

In this thesis, a support system for medical decisions is presented. It is dedicated to the diagnosis of a very serious respiratory disease caused by tobacco and named Chronic Obstructive Pulmonary Disease (or COPD). The system is based on the principles of case-based reasoning and will gather the experience of several experts doctors of the pulmonology department of the hospital Dorban (Annaba, Algeria).

An important question about the case base is that some attributes values are missing in most cases. In this context of missing data, four approaches with some variants have been proposed, implemented and evaluated. They aim to estimate the local similarity between two values when the at least one of them is missing, and they are executed during the *online* process of the system.

Two other approaches have also been proposed to fill the gaps left by the missing values with plausible values, using a statistical method and the principle of case-based reasoning itself.

Another contribution of this thesis is the modeling of the adaptation phase by an expert system. This integration is driven by the performance of expert systems to model the human reasoning that the expert arrives to explain. The main objective is to make the results of the most reliable system.

Keywords: Case-Based Reasoning, CBR, Rule-Based Reasoning, RBR, Diagnosis of the COPD, Medical Decision-making support System, Missing data