Table des matières

Liste des figures

Liste des tableaux

Introduction

Contexte de la thèse

Assurer une qualité de santé dans les établissements sanitaire est un élément essentiel de la compétitivité. Si les exigences en matière de temps et de qualité d'intervention sont les aspects les plus préoccupants, c'est la prévention et le diagnostic qui retiennent le plus souvent l'intention. Le personnel médical a des préoccupations d'offrir à leurs patients des soins de santé sécuritaire, d'où il doit intégrer des problèmes de sécurité, et d'efficacité du système de santé.

Cependant l'amélioration de la performance et des résultats du système de santé peut être possible seulement grâce à une gestion efficace des connaissances en matière de santé. Ces connaissances se basent, essentiellement, sur une bonne compréhension du dossier patient et du système d'information hospitalier (SIH).

Le dossier patient peut être défini de la manier suivante : « C'est l'ensemble des informations médicales, soignantes, sociales et administratives, qui permettent d'assurer la prise en charge harmonieuse et coordonnée d'un patient en termes de soins et de santé par les différents professionnels qui en assurent la prise en charge. C'est à partir du dossier patient que l'on assure la traçabilité de la démarche de prise en charge et c'est à partir de vues différentes des traces qu'il contient que l'on élabore des bilans d'activité et des travaux de recherche».

Les dossiers patients ont connus ces dernières années une évolution dans leur mode de gestion. Grâce à l'informatique, la possibilité de leur gestion a été améliorée. Ces dossiers patients sont collectés et utilisées pour différents objectifs. Ils sont enregistrés et utilisés principalement pour la gestion et l'analyse de l'état de santé de la population. L'existence de traces épidémiologiques précises permet le suivi et l'analyse des conditions sanitaires et sociales de la population. Certaines traces couvrent toute la population, et sont recueillis pendant des décennies. Ils sont fréquemment utilisés pour la recherche, l'évaluation, la planification et à d'autres fins par divers utilisateurs en termes d'analyse et de prévision de l'état de santé des individus.

Problématique

Une trace est un ensemble d'événements caractérisant le comportement d'un objet ou d'un individu pendant une tâche. Il existe plusieurs sortes de traces, par exemple une observation ou un enregistrement d'une interaction en vue d'une analyse. La trace d'interaction d'une personne pendant qu'elle réalise une tâche qui contient toutes les actions menées par l'individu pendant la tâche (e.g., tourner à droite, regarder dans le rétroviseur, etc.) ou la trace d'activité qui est un indice de l'activité des acteurs en précisant qu'il s'agit d'un résultat obtenu au cours ou au terme d'une activité, d'un événement ou d'un ensemble d'évènements.

Dans le domaine médical La collecte de l'information afin de construire une trace est très hétérogène. Une trace patient peut être appréhendée comme étant un ensemble d'évènements qui va subir le patient en raison de construire le dossier patient de ce dernier. Par exemple, les notes cliniques

et les interactions des médecins peuvent être vue comme étant des traces, alors la trace est un aide-mémoire pour les professionnels de santé. Une fois les traces médicales sont collectées il est important de pouvoir les exploiter afin d'identifier et d'expliquer des comportements anormaux, de les analyser pour construire un système d'aide à la décision clinique ou de définir des profils de tâches.

La fouille de données permet d'extraire automatiquement de l'information "pertinente" dans des masses de traces. Il existe plusieurs techniques de fouille, notamment des méthodes ensemblistes (règles d'association, analyse formelle de concepts) et des méthodes séquentielles (recherche de motifs séquentiels). Alors la fouille de données s'avère un moyen important pour le processus d'analyse des traces médicales ceci est reconnu sous l'acronyme fouille de données médicales FDM.

La FDM est généralement utilisée pour l'aide à la décision médicale et plus précisément pour l'aide au diagnostic médical afin d'offrir des systèmes capables de produire des décisions en temps réels aux médecins. Dans la FDM, la nature de traces diffère d'une spécialité médicale à l'autre. Dans notre cas d'études on s'intéresse aux maladies cardiovasculaires CAD qui ont été estimées comme étant la première cause de mortalité dans le monde. Nous souhaitons présenter un système d'aide à la décision clinique CDSS dédié aux traces CAD, tout en décrivant brièvement les approches de fouilles de traces utilisées pour sa mise en œuvre. L'objectif est double : d'une part proposer une approche de FDM, et d'une autre part l'intégrer dans un système d'inférence pour la prise de décision clinique. Dans cette thèse, notre objectif méthodologique consiste tout d'abord à expérimenter un ensemble d'outil d'extraction automatique des connaissances avec la fonctionnalité d'acquérir les connaissances existant dans les bases des traces médicales.

Contributions

L'objet de cette thèse est « fouille de données pour l'analyse des traces de patients ». Le lecteur de cette thèse devra prendre en considération le mot « analyse des traces » est indiqués dans le domaine médicale ce qui fait appel aux processus d'extraction de connaissances à partir de traces (ECT). Nos contributions s'articulent autour des principales étapes du processus ECT, dans un premier niveau, on propose une nouvelle approche de prétraitement de traces (première étape du processus ECT) et plus précisément une approche de sélection d'attributs (Mokeddem S et al, in press) Nous améliorons les performances des différentes techniques de classification en réduisant le nombre d'attributs impliqués dans la construction du modèle. Au lieu de faire participer l'ensemble de tous les attributs dans la classification qui entrainera une augmentation du temps de calcul, de l'espace mémoire et du bruit, nous proposons l'utilisation de notre techniques de sélection qui utilise l'algorithme génétique (GA) pour (i) la représentation des attributs, (ii) la génération des sous-ensembles d'attributs aléatoirement, et (iii) la technique Naïve bayes pour l'évaluation des sous-ensembles d'attributs générés. Ensuite, l'approche proposée a été appliquée

sur les traces médicales et plus précisément sur les traces des maladies cardiovasculaires (Mokeddem et al., 2014)(Mokeddem et al., 2016).

Dans la seconde contribution de la thèse, nous nous intéressons au raisonnement flou qui parait le plus adaptés au domaine médical grâce à sa capacité de représenté l'imprécision et l'incertitude. Nous proposons cette fois-ci une approche pour désigner un système flou d'aide au diagnostic médicale. L'originalité de ce travail réside au fait que notre approche est basée dans un premier temps sur (i) une technique d'extraction de règles automatique pour (ii) les intégrés dans une base de connaissance floue afin d'avoir (iii) la meilleur modélisation floue des traces (Mokeddem and Atmani, 2016).

Notre motivation dans le cadre de notre thèse réside dans le fait qu'un tel système offre une amélioration de la qualité de soins donc il faut assurer la qualité d'intervention au profit du patient et ceci en fournissant :

- L'accès en ligne à des informations de référence dans le contexte d'une situation clinique donnée
- La recherche et présentation des traces cliniques pertinentes dans le contexte de la tâche en cours : décision diagnostique ou thérapeutique, prescription médicamenteuse, etc.
- L'aide à la documentation des soins sous la forme de listes de traces cliniques pertinentes :
  - Recueillies afin d'établir un diagnostic ou un pronostic ou de suivre les effets d'un traitement,
  - Associées à des contrôles automatiques de la qualité des traces saisies
- L'aide à la prescription des actes diagnostiques ou des médicaments au moyen de formulaires établis à partir des recommandations de pratiques et proposant des bilans ou protocoles appropriés à la situation clinique du patient.
- Les fonctions de gestion de protocoles pour la prise en charge de maladies chroniques utilisant les diverses modalités d'intervention.
- Les alertes informant les cliniciens de la survenue d'évènements, tels que l'identification d'un résultat d'examen anormal, la détection d'une allergie ou d'une interaction médicamenteuse dangereuse.
- Les rappels ou « aide-mémoire» rappelant à l'utilisateur :
  - Soit des recommandations pour la prévention primaire ou secondaire,
  - Soit des recommandations pour le diagnostic, la prescription d'examens ou de médicaments, la surveillance d'un traitement.

Organisation de la thèse

La thèse est structurée en six chapitres. Le chapitre 1 intitulé, « Activités, Traces et connaissances », présente le cadre de cette thèse. Nous positionnons face aux questions : qu'est-ce qu'une activité ? Qu'est-ce qu'une trace ? Comment collecter les traces ? Nous verrons que l'analyse de ces traces nécessite un travail d'interprétation, et donc l'intervention de techniques d'analyses issues de l'intelligence artificielle et fouille de traces.

Le chapitre 2 intitulé, « Fouille de Données pour l'Analyse des Traces Patients », introduit les différents de concepts utilisés dans l'analyse et la fouille des traces patients, le raisonnement utilisé et le domaine médical choisit dans ce processus. Les systèmes d'aide à la décision sont aussi introduits, dans ce même chapitre, à cet effet, une panoplie d'outils dédiés à l'analyse des traces est succinctement décrite.

Le chapitre 3 intitulé, « Analyse de la Pertinence des Variables pour la Réduction de Dimension pour la classification », montre la première contribution de la thèse, ce chapitre est scindé en deux parties. La première donne un bref aperçu sur les techniques de prétraitement de traces, l'importance de la pertinence des traces dans un processus de la fouille des traces. Par la suite, la technique de sélection des descripteurs pertinents des traces proposée est décrite dans la deuxième partie.

Le chapitre 4 intitulé, « Une Méthode Effective pour la Sélection d'Attribut basée Algorithme Génétique Wrapper Naïve Bayes », évalue notre première contribution, ce chapitre est scindé en deux parties. La première donne un bref aperçu sur les différents bases de traces utilisées et évalue notre contribution en les utilisant. Dans la deuxième partie, on montre la performance de notre contribution pour la sélection des descripteurs de traces pertinents.

Le chapitre 5 intitulé, « Un CDSS Basé Logique Floue pour l'Aide au Diagnostic des CADs » donne un bref aperçu sur la modélisation et le raisonnement flou des traces dans la conception les systèmes basé connaissances en montrant leur intérêt pour l'analyse des traces. On présente notre contribution pour concevoir un CDSS pour le CAD. Enfin, on illustre la phase de l'évaluation des décisions produites par notre système en matière de performance et de précision en temps et de qualité de décisions. Ce chapitre présente la problématique et notre contribution flou, l'architecture fonctionnelle de notre CDSS, l'identification des critères d'évaluation, l'élaboration des décisions et leurs évaluations.

Cette thèse s'achève par la présentation d'une conclusion synthétisant les différents résultats obtenus et ouvrant des perspectives à la présente étude.

Chapitre 1

Activités, Traces et Connaissances

Sommaire

1.1 Introduction

L'idée de conserver un enregistrement d'un historique des opérations effectuées par une machine au cours de son utilisation, ou un grand nombre d'instructions est exécutées, afin de comprendre les résultats et vérifier les étapes de ce calcul est devenue primordial avec l'avancée de l'informatique.

Ce principe, qui a l'origine était utilisé pour comprendre les instructions effectuées par la machine, a tout simplement été appliqué pour enregistrer et comprendre les transactions et les données d'un système informatique. Ceci correspond généralement à l'arrivée des systèmes informatiques avec des bases de données et des interfaces graphiques et leurs utilisation croissante dans tout les autres domaines. Cette évolution dans ces systèmes a permis de modéliser n'importe quelle « activité » de la vie réelle et l'erengistrer sous forme de « trace ». Par la suite cette démarche d'enregistrement des traces, facilitée par les progrès techniques réalisés sur les environnements numériques (rapidité, capacité de stockage, etc.), a été appliquée dans de multiples domaines de recherche : Informatique bien sûr, mais également Ergonomie, Psychologie Cognitive, Sciences de l'Information. Ces traces permettent en effet d'analyser en détail ces activités au-delà de l'imminence spatiale et temporelle d'une observation.

Dans le monde de l'analyse des traces on trouve souvent les mots « activité », « trace » et « connaissance ». Une activité constitue généralement un processus, ou ensemble d'actions tant dis qu'une trace restitue ce qu'on observe de ce processus. Lorsqu'on on introduit l'analyse et l'interprétation des traces, on fait appel à la notion de connaissance et plus précisément aux outils d'extraction de connaissances. Dans ce chapitre, nous visons à introduire le sens des termes activité, trace et connaissance. Nous montrons aussi la relation entre ces termes et nous donnons le positionnement scientifique à l'égard de notre thèse

1.2 Activité

Le dictionnaire Larousse présente le mot « activité » en quatre acceptations : la première est associée à un ensemble de phénomènes, la deuxième est par rapport à une faculté, la troisième relevant une action et la dernière relative à un ensemble d'actions, dans notre cas on s'intéresse à la troisième :

> "Action de quelqu'un, d'une entreprise, d'un pays dans un domaine défini ; champ d'action : Activités professionnelles. Une usine qui étend son activité à de nouveaux secteurs. Ensemble des actions diverses menées dans un secteur, ou qui se manifestent dans un lieu : Une période d'intense activité diplomatique. (Larousse, 2013)"

L'activité se définit intuitivement par « ce qui se fait dans une situation particulière Rabardel and Samurçay (2001). En effet, une activité est un ensemble « d'actions », de « phénomènes

manifestant un processus ». Dans son acception large, l'activité est aussi bien produite par des organismes vivants, humains, animaux, que par des processus physiques ou chimiques : activité d'un volcan, activité du soleil. Par définition, toute activité se déroule au cours du temps, c'est un processus dynamique.

L'activité relève de processus unifiés, c'est « l'ensemble des actes coordonnés des travaux de l'être humain » (Larousse, 2013), et vise à atteindre un but. L'activité d'un patient, par exemple, est l'ensemble de ses bilans et bulletins clinique, consultations, maladies atteintes, etc. Pour atteindre un but : la guérison et le bienêtre. Ainsi, l'activité est caractérisée par un ensemble de processus unifiés (Leont'ev et al., 1984) visant à la réalisation d'un objectif global, commun à l'ensemble de ces processus. Du point de vue l'état de santé du patient, l'environnement lui-même s'apparait comme un environnement dynamique, c'est-à-dire évolué par lui-même sans intervention de sa part.

Une activité, à travers son objectif, s'intéresse à changer l'état de quelque chose. Sans environnement où procéder, il n'y a pas d'activité. L'activité est donc inséparable du contexte dans lequel elle se réalise. Ce que l'on peut constater de l'activité, ce sont que les actions observables, c'est-à-dire que la partie visible des processus mis en œuvre pour réaliser l'activité. Ces actions observables, ce sont les comportements. Néanmoins, l'activité ne se limite pas aux comportements, car elle contient aussi les processus qui procréent le comportement. Par exemple, le raisonnement auquel le médecin suit pour faire un diagnostic fait partie de son activité. Toutefois, ce choix lui-même n'est pas observable. Ce qui est observable, ce sont les actions qu'il met en œuvre pour diagnostiquer : vérifications des symptômes, prescriptions des médicaments, etc.

Dans la suite de notre thèse on retient la définition suivante : L'activité est un ensemble d'actions observables ou non, introduites dans une situation et visant à la réalisation d'un objectif global.

1.3 Outils pour étudier l'activité

L'homme a probablement toujours porté une réflexion sur son activité pour tenter de l'optimiser. Cette réflexion sur l'activité dans le but de l'optimiser s'est développée en science avec la création de la science d'ergonomie. Cependant, l'ergonomie est « l'étude scientifique de la relation entre l'homme et ses moyens, méthodes et milieux de travail et l'application de ces connaissances à la conception de systèmes qui puissent être utilisés avec le maximum de confort, de sécurité et d'efficacité par le plus grand nombre. » (Amalberti, 1996).

Pour étudier l'activité, l'ergonomie dispose d'outils et de méthodes (Gillet, 1987). Nous en présentons ici quelques exemples caractéristiques : l'observation, l'expérimentation, les entretiens et les enquêtes.

1.3.1 Observation

L'observation de l'activité constitue la base pour étudier une activité. Le but est de comprendre l'activité, dans sa complexité naturelle, c'est-à-dire telle qu'elle se voit naturellement. L'observation consiste à observer l'opérateur en situation, dans son environnement naturel. « Ce qu'on observe, ce sont des traces (des actions, des verbalisations, des résultats d'activité) qui serviront à inférer l'inobservable. » (Gillet, 1987).

1.3.2 Expérimentation

Par contre à l'observation qui s'intéresse plutôt à interpréter les faits dans leur complexité et à les confronter à un cadre théorique, l'expérimentation vise à tester une hypothèse en soumettant l'opérateur à la réalisation d'une tâche. L'expérimentation cherche à reproduire les conditions de l'activité en laboratoire pour tester. L'expérimentation permet un meilleur contrôle, mais s'écarte de la situation réelle étudiée surtout dans les environnements simulés.

1.3.3 Entretiens

Les entretiens avec les opérateurs sont nécessaires pour analyser les formes cognitives de l'activité. Cependant, comme nous l'avons évoqué précédemment, la tâche telle que l'opérateur se la définit et l'activité telle qu'il l'exécute peuvent être différentes. L'opérateur n'a pas forcément conscience de cette différence. Ainsi, pour éviter que l'opérateur ne rationalise a posteriori un comportement qui ne l'est pas forcément au moment où il engage son activité ou dont il n'avait pas forcément conscience, il est nécessaire d'utiliser des techniques d'entretien semi-supervisés, comme l'entretien d'explicitation (Vermersch, 1994).

1.3.4 Enquêtes

Les enquêtes et les questionnaires sont des techniques exploratoires qui permettent de guider vers les éléments de l'activité à étudier plus en détail.

On peut dire que l'activité est un processus complexe qu'il est nécessaire d'étudier et d'analyser dans sa complexité et en condition naturelle en prenant en compte la dynamique de la situation. Pour interpréter l'activité dans sa totalité, il est nécessaire de s'appuyer sur des théories, issues des sciences cognitive, et de les combiner à des techniques d'observation et d'entretien. Les observations permettent la récolte des traces avec lesquelles on pourra déduire les éléments inobservables de l'activité.

1.4 Traces

Le mot traces a plusieurs acceptations selon le dictionnaire de Larousse. la première relève de l'empreinte et de l'action, la troisième est relative au littéraire, quatrième et cinquième sont spécifique aux mathématiques et psychologie (Larousse, 2013):

1. Suite d'empreintes laissées sur le sol par le passage de quelqu'un, d'un animal, d'un véhicule : Relever des traces de pas dans une allée. Marque laissée par une action quelconque : La porte garde des traces d'effraction. Très faible quantité d'une substance : Déceler des traces d'albumine dans les urines.
2. Littéraire. Ce qui subsiste de quelque chose du passé sous la forme de débris, de vestiges, etc. : Des traces d'une civilisation très ancienne.
3. Mathématiques : En géométrie descriptive, intersection d'une droite ou d'un plan avec l'un des plans de projection. Psychologie : Ce qui subsiste dans la mémoire d'un événement passé.

Pour illustrer ce qu'est une trace, nous allons nous appuyer sur des exemples médicaux. Ces exemples, bien que simple, nous permettrons de montrer les questionnements liés à l'utilisation de traces : que sont les traces ? Quelles sont leurs significations ? Que permettent-elles de déduire ? Dans quelle mesure ?

Comme le soulignent les définitions du dictionnaire, une trace est représentée par « l'activité d'un objet ou d'une personne en laissant une empreinte » donc une trace résulte alors d'une activité. Par exemple dans le domaine médical, on trouve les traces de soins, qui ont comme rôle, l'identification du patient et les professionnels qui ont édicté les soins au patient en retrouvant toutes les informations concernant le processus de soins promulgués au patient. Donc on peut retenir la définition suivante :« Une trace est un ensemble d'événements caractérisant le comportement d'un objet ou d'un individu pendant une activité » . Une trace permet typiquement d'apprendre des choses sur le phénomène, l'individu, etc...qui l'a laissé. Dans notre exemple, si on a la trace de soins, on pourra connaitre le diagnostic du médecin, les symptômes du patient, réaction du patient envers le traitement, et plus général l'état du patient pendant et après les soins.

Les « traces numériques » font de plus en plus souvent leur apparition dans l'actualité (dans la suite on utilise le terme trace pour référer a une trace numérique). Lorsqu'elles sont évoquées il s'agit souvent de désigner des moyens de « traçabilité ». Dans le monde informatique, la trace numérique est une trace construite autour d'empreintes numériques laissées volontairement (ou non ?) dans un système informatique. En effet, cette notion désigne les informations qu'un dispositif numérique enregistre sur l'activité ou l'identité de ses utilisateurs, soit automatiquement,

soit par le biais d'un entrepôts de données. Moteurs de recherche, blogs, réseaux sociaux, sites de commerce électronique, Électrocardiogramme, SIH, dossier patient et tous systèmes qui requièrent une identification ou une interaction sont susceptibles de capter des informations sur l'utilisateur, parcours, requêtes, préférences, achats, connexions, évaluations, coordonnées, bilans médicaux, diagnostic.

Le monde numérique est donc potentiellement normalisateur de la production de traces à partir d'empreintes plus ou moins contrôlées dans leurs inscriptions. En se basant sur la notion d'empreintes numériques pour construire une trace informatique, plusieurs constats relatives au caractère numérique de la traces peuvent être présentés:

- Introduire l'empreinte numérique pour construire les traces implique un codage de ces empreintes dans le système informatique.
- Le système informatique doit modéliser les traces numériques dans un format cohérent.
- Le système informatique doit géré les diapositives liés à l'inscription des empreintes numériques et doit fournir des outils d'analyse des traces.
- Il est toujours possible de faire de nouvelles traces numériques avec des empreintes numériques existantes.

Les traces sont donc des données regroupées, traitées et combinées dans d'importantes bases de données. L'analyse de ces traces peut révéler des informations significatives, stratégiques ou sensibles. Plusieurs applications de l'analyse des traces existes, par exemple, Les traces numériques peuvent en particulier être utilisées pour profiler les patients, par extraction automatique d'un profil à partir de l'observation de leurs dossier patient. Ce profilage peut servir ensuite à étudier l'état de santé de la population.

1.4.1 Interprétation et collecte des traces

Une trace fait l'objet d'interprétations et d'analyse. Pour inférer toutes ces informations sur le patient et son état de santé, les analystes sont amenés à analyser ces traces. L'analyste va chercher dans la trace les indices nécessaires à son interprétation. Mais cette interprétation ne pourrait se faire sans des connaissances supplémentaires dont dispose l'analyste : sur le domaine médical, par exemple les maladies et leurs diagnostic, sur le contexte, par exemple les conditions sociaux économique du patient, ou sur les traces elles-mêmes, par exemple comment l'état de santé du patient évoluent avec le temps?

Lorsque l'analyste observe les traces d'un patient, il va en déduire des informations sur le patient. À partir exactement des mêmes traces, un médecin serait susceptible d'ignorer certaines observations relatives au patient pour s'intéresser au fascinant phénomène de réaction et de l'évolution de la forme des cellules sanguines au fur et à mesure de son traitement. Pourquoi

le médecin ignorerait-il ces observations ? Pourquoi l'analyste ignorerait-il la forme des cellules ? Par conséquent, les informations décrivant la trace est ceux que l'on décide d'observer. C'est l'observation, c'est-à-dire la collecte de la trace, qui donne le statut de trace aux éléments observés.

Une trace est issue d'une observation nous appelons cette observation qui crée la trace la « collecte de la trace ». Dans notre exemple, la trace est laissée par le patient et est collectée a posteriori par le personnel médical. Qu'en est-il de l'analyste qui installe une caméra pour enregistrer le comportement des patients ? On pourrait penser qu'il y a une différence de nature entre une trace fortuite laissée par l'état de santé du patient et une trace enregistrée volontairement par un dispositif technique. Cette question nous relève trois dimensions de la trace : 1) la construction d'observables, 2) l'observation de ces observables, qui crée la trace et enfin 3) l'interprétation et l'analyse de la trace.

Une observation est partielle et engagée. Une trace est la trace de ce que l'on décide subjectivement d'observer. Elle est le résultat d'une observation subjective, qui répond au point de vue de l'observateur. Lorsque la trace est analysée pour étudier scientifiquement des activités ou bien des objets, l'observation fait l'objet de beaucoup de précautions pour permettre son objectivation et la construction de connaissances valides sur l'activité étudiée.

Les interprétations relèvent d'un point de vue. À partir de la même trace, c'est-à-dire des mêmes « données », les interprétations peuvent diverger. Tout d'abord, nous l'avons souligné, la même trace va produire des interprétations de nature différente en fonction de qui la regarde (le médecin ou l'analyste). Mais même lorsque deux médecins observent les mêmes traces, leur interprétation peut diverger. Ce n'est pas étonnant puisqu'il s'agit en réalité de déduire ce qui n'a pas été observé à l'aide des simples indices présents dans une trace. Ces indices ne sont pas toujours satisfaits, surtout lorsque l'on cherche à déduire des éléments, par nature non observables, comme le raisonnement du médecin afin d'inférer son diagnostic. Notons que la situation est différente dans le cadre de l'interprétation scientifique des traces.

En résumé on peut dire qu'une trace est produite par une activité, mais c'est l'observation (la collecte) qui est l'élément créateur de la trace. Une trace contient des éléments relatifs aux actions, et d'autres aussi à leur contexte. Ces indices sur l'action et sur son contexte permettent de déduire ce qui n'a pas été observé ou ce qui n'est pas observable (Analyse de traces).

1.4.2 Analyse des traces

Les activités et les traces sont liées. Nous avons montrés que le processus principal pour étudier l'activité est de collecter les traces en se basant sur les observations. L'interprétation des traces permet d'analyser l'activité en se basant sur le modèle de la trace qui le représente. Par ailleurs, Les traces, sont des traces d'actions dans un contexte et contiennent des indices de l'activité. Étudier les traces permet donc d'appréhender l'activité et l'état de l'objet qui l'a produit. Anal-

yser des traces, c'est interpréter les indices qui y sont présents pour leur donner du sens. Cette interprétation permet de construire des connaissances sur l'activité étudiée. La figure 2.7 résume l'ensemble du processus.

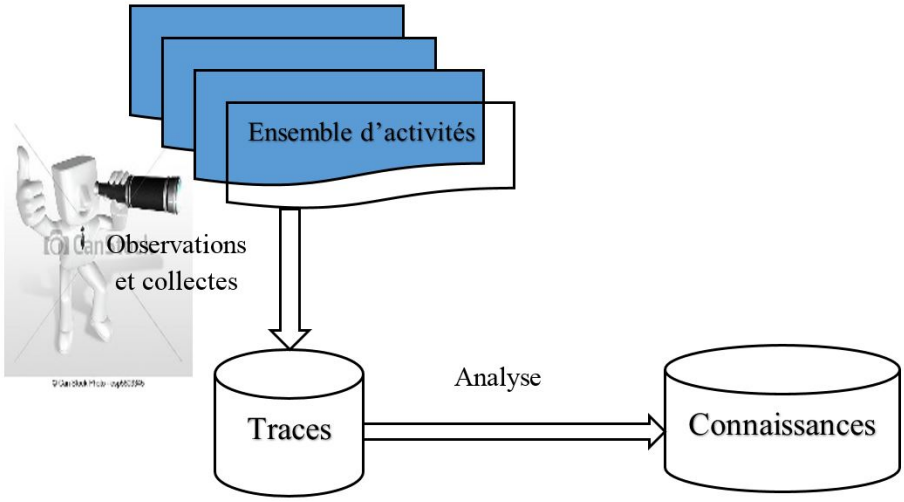


Figure 1.1: Démarche générale pour l'analyse des traces

Cependant, nous n'avons pas encore précisé quel est le statut de la connaissance. la section suivante réponds aux questions relatives à l'ingénierie des connaissances : Qu'est-ce qu'une connaissance ? Au-delà des connaissances sur l'activité, peut-on construire des modèles de l'activité ? Comment, à partir de l'analyse du modèles des traces, on extrait des connaissances ? À l'issue de cette section, nous détaillerons en quoi ce travail de thèse éclaire ces différentes questions.

1.5 Connaissance

Lorsque nous disons que l'analyse des traces permet à celui qui observe la trace de connaître l'activité, nous introduisons le concept de connaissances. La connaissance à laquelle nous nous intéressons est la connaissance relative à la trace. Nous pouvons distinguer plusieurs types de connaissances, par exemple les connaissances que le médecin peu savoir sur ses patients et sa propre activité.

1.5.1 Connaissances sur l'activité

En ingénierie des connaissances, Une connaissance, comme le souligne Bachimont (Bachimont, 2004a), est la « capacité à réaliser une action pour atteindre un but ». Dans ce contexte, la trace contient des informations sur les activités de l'opérateur (Médecin) qui nous intéressent, car c'est lui qui met en œuvre l'activité. Les connaissances de l'opérateur sont en partie explicites et en partie implicites. Les connaissances explicites sont des connaissances clairement articulées dans

différents supports, document, systèmes informatiques, etc., et que lorsque l'opérateur sait comment les exprimées. C'est le cas des connaissances scientifiques par exemple. Ces connaissances sont facilement transférables car elles apparaissent sous une forme tangible. Les connaissances explicites sont stockées ou diffusées sur des supports d'information matériels.

Les connaissances sont implicites quand il s'agit du « savoir faire » ou l'opérateur ne peut pas les exprimées et expliquées (Anderson, 2013). L'accumulation des connaissances implicites se fait par la tradition et le partage d'expériences notamment aussi par apprentissage. Les connaissances de l'opérateur sur l'activité ne sont pas accessible juste par connaissances explicites. Certaines connaissances sont lui accessible lors de l'exécution de l'activité. Autrement dit, l'opérateur détient des connaissances pour la réalisation de l'activité sans avoir la conscience de possédées ces connaissances. Cependant, les connaissances construites relativement a l'activité ne sont pas nécessairement les connaissances permettant l'exécution de cette dernière. Par conséquent, les connaissances sur l'activité sont indépendantes et différentes de celle utilisé pour la mise en œuvre par l'opérateur pour réaliser l'activité.

l'analyse des activités permet de reconstruire les connaissances de l'opérateur implicites et explicites par le bias des observations et entretiens avec l'opérateur. (Georgeon, 2008) proposait une approche pour analyser la partie observable de l'activité, en se basant sur cette approche, nous nous cherchons a exploiter les traces relatives a l'activité pour modéliser la partie cacher de l'activité: les connaissances que l'opérateur utilise pour la réalisation de l'activité.

Dans la suite de notre thèse, nous appellerons analyste la personne qui observe et analyse les traces d'activité. Les activités étudiés sera l'activité médicale appliquée sur les patients du service cardiologie. L'objectif de l'analyste est de comprendre l'activité humaine dans sa complexité naturelle en extractant des connaissances pour améliorer l'activité étudiée (la rendre plus efficiente, réduire les risques, et proposer des conseils de préventions, etc.).

Les connaissances construites par l'analyste représentent l'état de santé des patients et sera similaire aux connaissances des experts. L'objectif est donc de modéliser les connaissances extraites à partir de traces collectées dans un système d'aide à la décision. Dans cette thèse, les connaissances relèvent de deux niveaux : les traces qu'a l'analyste doit collecter et les modèles que construit l'analyste pour extraire les connaissances. Notons ici que l'analyste est aussi un opérateur (Falzon, 1991).

L'activité qu'il exécute est une activité d'analyse et de modélisation. L'analyste exploite des connaissances qui sont relatives à son activité (d'analyse). Construction des connaissances L'observation est une source de connaissances. Mais quel est le lien entre l'analyste, les traces qu'il interprète et les connaissances ? L'ingénierie des connaissances et l'épistémologie donnent un cadre à ce problème.

Une trace n'est pas en soi une connaissance. C'est l'analyse et l'interprétation de la trace qui constitue une connaissance. Ainsi, en reprenant l'expression de Bachimont (Bachimont, 2004b), la trace serait une inscription de connaissances, du fait qu'elle est sujette à une analyse et inter-

prétation et que c'est cette interprétation de la trace qui est connaissance : « les connaissances ne s'appréhendent qu'à travers des inscriptions matérielles qui les expriment, et dont elles sont l'interprétation. ».

Notons ici que la connaissance issue de l'interprétation de la trace n'est pas nécessairement une connaissance pour l'opérateur, mais plutôt une connaissance pour l'analyste. À partir de ces connaissances d'analyse, l'analyste déduit les connaissances de diagnostic qu'il estime être celles de l'opérateur. Ainsi, la construction de sens se fait toujours par l'humain. Cette interprétation s'appuie non seulement sur des éléments matériels, tels que les traces, mais aussi sur d'autres connaissances que possède l'analyste. Ces autres connaissances sur le domaine étudié ou sur l'activité étudiée, dont nous présenterons un aperçu au chapitre 2, permettent à l'analyste de construire l'interprétation de ce qu'il observe.

1.5.2 Connaissances et Modèles

Qu'est-ce qu'un modèle ? Leplat nous éclaire sur la question (Leplat, 2003). Il cite deux définitions que nous reprenons ici. La première définition, de Régnier, est la suivante : « Un modèle d'un objet concret est un objet abstrait dont la description est tenue pour nous pour une description dudit objet concret. » La seconde définition formulée par de Montmollin présente un modèle comme « une représentation des comportements d'opérateurs dans une situation de travail ou de vie, et permettant d'agir sur cette situation ».

Leplat désigne la différence entre le modèle et l'objet a modélisé, Cette distinction souligne l'importance qu'un modèle doit être une simplification de l'objet lui même (activité) (Leplat, 2003). Un modèle tend vers une caractérisation de certains traits de l'activité modélisé. En addition, un modèle vise aussi « à définir leurs relations afin de mieux comprendre le fonctionnement de ce système et de donner les moyens de le modifier » (Leplat, 2003).

En effet, selon l'objectif que le modèle doit répondre, on peut classer les modèles en trois types: modèles descriptifs, modèles explicatifs ou bien modèles prédictifs (Sperandio, 2003). Un modèle descriptif de l'activité permet de décrire ce l'activité. Un modèle explicatif permet de comprendre les raisons sous-jacentes qui ont engendré cette activité. Un modèle prédictif permet de prédire le comportement futur de l'opérateur. Un modèle peut posséder une ou plusieurs de ces facettes. Particulièrement, l'ergonomie considère le caractère explicatif du modèle afin d'expliquer l'activité. Dans le cas idéal, un modèle est caractérisé par les trois types: descriptif, explicatif et prédictif. On s'intéresse beaucoup plus au modèle explicatif et prédictif dans la suite de la thèse.

Une trace n'est que l'inscription d'une histoire produite par l'activité. Par nature, une trace contient donc des informations descriptives sur l'activité. C'est seulement par l'interprétation que l'analyste peut formuler des hypothèses permettant d'expliquer les comportements observés. Un modèle explicatif nécessite donc l'intervention de l'analyste qui va se servir de ses connaissances

sur l'humain et sur l'activité pour produire ces explications.

Contrairement au modèle explicatif, une explication relève généralement d'une connaissance élémentaire sur l'activité. Un modèle explicatif est une représentation structurée et intégrée dans un contexte plus général de l'activité. Un modèle descriptif donne des informations sur l'activité dans son instanciation, c'est-à-dire sur des instances de l'activité. Par conséquent, les traces peuvent être appréhendé comme étant une description de l'activité. Contrairement, la structuration des connaissances produite par un modèle explicatif fournit des informations sur l'activité. Neisser (Neisser, 1976), puis (Stanton et al., 2009) ont construit un modèle explicatif en se basant sur les deux concepts: phénotypes[1] et le génotype[2] empruntée au domaine de la génétique. Le génotype exprime une connaissance sur l'activité dans ses potentialités, tandis que le phénotype correspond à une expression effective de ce génotype au moment ou l'activité est réalisée. Autrement dit, les traces représentent la partie phénotype de l'activité c'est-à-dire l'activité telle qu'elle est exprimée, alors que le génotype représente le modèle de l'activité.

Une activité implique l'humain dans différent aspects, par exemple dans une activité médicale on trouve les aspects: sociaux, cognitifs ou biologiques. Dans l'aspect social, l'opérateur peut avoir une certaine vision de son environnement et de ses comportements vis a vis la société (acceptable ou pas par les lois, un risque, etc.). Pour l'aspect cognitif, l'opérateur utilise un raisonnement durant l'activité afin de prendre des décisions ou il peut inclure sa perception dans l'environnement regardant sa situation (représentation mentales), ses compétences, savoir-faire, il implique aussi son mémoire (connaissances explicites), etc.. Dans l'aspect biomédicales, l'opérateur fait appel a des outils et analyse le résultat de obtenue avec chacun de ses outils, comme des analyse biologique, imagerie médicale ....

L'activité est un processus unifié ou l'opérateur n'est pas scindé selon chacune des aspects sociales, cognitives, biologiques lorsqu'il réalise une activité. En effet, un modèle d'activité doit donc considérer les éléments de l'opérateur qui sont pertinents en prenant compte des objectifs de l'activité et faire en sorte de proposer une vision unifiée de ces éléments. Dans le cadre de médicale, par exemple, l'aspect social de l'activité permet d'étudier les comportements de certaines populations vis-à-vis de la prise de risque d'atteindre des maladies (Banet, 2010), alors que

[1]En génétique, le phénotype est l'ensemble des caractères observables d'un individu. Très souvent, l'usage de ce terme est plus restrictif : le phénotype est alors considéré au niveau d'un seul caractère, à l'échelle cellulaire ou encore moléculaire. L'ensemble des phénotypes observables chez un individu donné est parfois appelé le phénome. Le concept de phénotype est défini par opposition au génotype, l'identité des allèles qui caractérise le génome d'un individu. Pour certains traits simples, la correspondance entre le génotype et le phénotype est directe, et les deux sources d'information sont redondantes. Cependant, la plupart des caractères (les caractères qualitatifs) dépendent de multiples gènes, et l'influence du milieu (l'environnement dans lequel l'organisme se développe et vit) peut être un facteur déterminant. Dans ce cas, le génotype ne permet pas de prévoir précisément le phénotype de l'individu, mais seulement d'estimer sa valeur moyenne.[Wikipedia]

[2]Le génotype est l'ensemble des caractéristiques génétiques d'un individu. Les gènes d'un être vivant se retrouvent normalement chez tous les membres de son espèce. Toutefois, il existe plusieurs versions de chaque gène. On appelle ces versions différentes d'un même gène des allèles. Cela donne des possibilités de combinaison différentes entre les gènes extrêmement élevées. Chaque génotype est ainsi habituellement unique, sauf cas exceptionnels comme la gémellité monozygote (« vrais jumeaux »). On désigne ainsi souvent familièrement le génotype comme la « carte d'identité génétique » d'un organisme. [Wikipedia]

l'aspect biologique permet d'améliorer l'état de santé de la population par voix de recommandation et réduire la gravité d'atteindre ces maladies, en étudiant les trace typique et pertinentes des activités des médecins dans des situations différentes (Hétier, 2009). Il est impossible de construire un modèle complet de l'humain décrivant toutes les dimensions de l'activité.

Nous avons démontrés qu'une trace est un ensemble d'informations liées à l'activité telle qu'elle s'est indiqué dans un domaine donné. Ces informations partielles ne constituent que la partie mesurable de l'activité, par ailleurs, l'activité ne se réduit pas par les ensembles mesurables de l'activité mais c'est un atout qui tend à une situation unique avec des objectifs communs. Cependant, un modèle de l'activité doit être unifié et construit en analysant les traces d'activité donc faire en sorte d'identifier les unités de l'activité pertinents à travers la lecture morcelée qu'en donnent les traces.

Nous l'avons décrit plus haut que les modèles explicatifs évoque l'activité en intension, tandis que les traces représentent l'activité en extension. Il existe donc une relation épistémologique entre activité et traces. Cette relation est étudiée en informatique ou il est possible de construire des modèles à partir d'un ensemble de traces en les analysants (Diekert et al., 1995). La correspondance entre modèle synthétisant un ensemble de traces et les traces que ce modèle peuvent concevoir sont particulièrement étudiées dans le cadre de modèles automates tels que les automates à états finis ou réseaux de Petri. Dans ce contexte, la qualité du modèle construit à partir de traces se traduit ainsi : « considérant une activité produite par un processus dont le modèle peut être formaliser par le formalisme A„ en usant le formalisme A, peut-on générer un modèle de l'activité, à partir des traces de cette activité? ».

Cette question pose deux problèmes. Le premier problème est qu'on ne peut pas connaitre a priori si le processus de construction de l'activité peut se modéliser avec le formalisme A. Si l'objectif est de rechercher a trouver un modèle exact, il est impossible de savoir a priori si le formalisme A permet de modéliser exactement le processus étudié. Si par contre on s'autorise un modèle simplificateur de l'activité, il est possible de choisir le formalisme A en fonction de sa capacité à traduire les traits caractéristiques de l'activité étudiée. Le deuxième problème est empirique. La modélisation à partir de traces relève des critères sur l'exhaustivité de la trace et la présence de suffisamment d'informations pour produire le modèle. Dans le cas de traces d'activité médicale, les traces ne sont pas exhaustives ou souvent on trouve des informations manquantes: elles sont partielle et certaines informations nécessaires à la compréhension de l'activité manquent car la trace c'est ce qu'on décide d'observer lors de l'exécution de l'activité.

Dans cette thèse, nous nous focalisons à la fois sur l'étude des traces médicale et sur celle de la performance qui en résulte. Comment construire un modèle d'activité ? Un modèle d'activité se construit en structurant les connaissances issues de l'observation de l'activité, c'est-à-dire des traces d'activité. Cette structuration des connaissances sur l'activité s'effectue, lorsque cela est possible, en accord avec d'autres modèles (et théories) issus de disciplines telles que les sciences cognitives. La boucle complète de construction de modèles de l'activité est présentée à la Figure

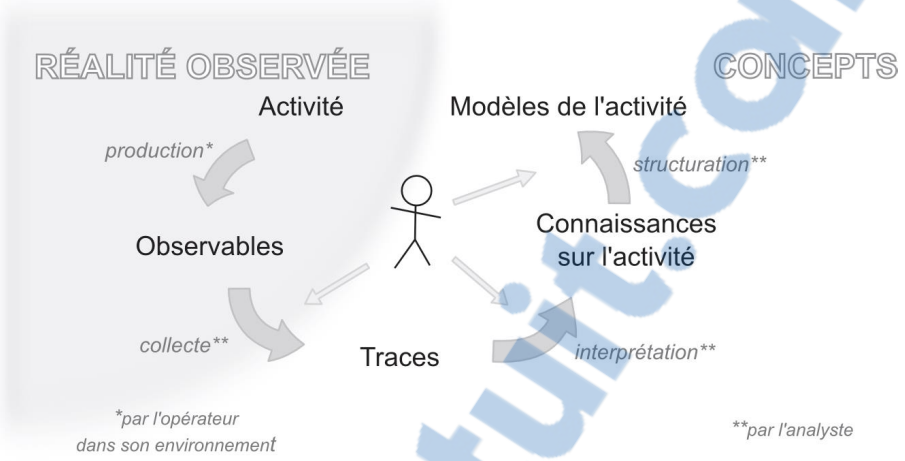


Figure 1.2: Analyse et modélisation de traces.(Georgeon, 2008)

1.2. Cette figure souligne le rôle de l'analyste dans la construction des modèles de l'activité.

L'objectif de cette thèse est de permettre la construction de modèles à partir de traces. Dans notre cas la collecte de la trace se fera à partir des données médicales. Il est donc nécessaire de s'attarder plus particulièrement sur la question de l'interprétation des traces et de la construction de modèles à partir de traces.

1.6 Outils pour modéliser les traces

Ce chapitre a permis de poser les questions relatifs à notre travail de thèse. Ce contexte permet de situer notre travail de thèse vis-à-vis ces questions qui englobent plusieurs disciplines. Au-delà des questions déjà soulevées, l'articulation entre sciences cognitives, ingénierie des connaissances a été évoquée. Par conséquent, il est à présent possible de présenter plus clairement les objectifs de cette thèse. Cette thèse vise à concevoir des outils pour analyser les traces, lui permettant d'assister l'opérateur durant son activité en se basant sur des connaissances extraites à partir des traces.

L'activité est un processus unifié, visant un but. Une activité humaine prend place au sein d'une situation. L'activité implique des composantes non observables, telles que la connaissances du sujet et les composantes observables, produits de l'activité au sein de la spécificité de la situation. Les éléments observables de l'activité, tels qu'ils sont collectés par un observateur, constituent des traces de l'activité. Ces traces sont le matériau à partir duquel nous cherchons à reconstruire des modèles de l'activité. Cependant, les traces ne sont que des inscriptions de connaissance. Pour produire des connaissances sur l'activité, il est nécessaire d'interpréter les traces. Pour produire des connaissances sur les aspects non observables de l'activité, il faut aussi effectuer des inférences à partir des traces.

L'opérateur, ici, est un expert médical, car il étudie l'activité du patient et son état. Dans cette thèse, l'activité étudiée est l'activité du patient, dans ses composantes comportementales et cognitives. Le but du modèle est de mieux comprendre l'activité. Ce modèle doit donc être un modèle en intention, plutôt que modèle en extension (En définissant les caractéristiques du modèle). Ainsi, le modèle que l'on cherche à construire dans cette thèse doivent synthétiser l'activité. Il s'agit de construire de modèle explicatif, éventuellement prédictif, de l'activité.

Un système à base de traces (SBT) est "un système qui enregistre systématiquement les interactions d'un utilisateur avec un système informatique, tout en s'intéressant à la finalité de ces enregistrements" (Laflaquière, 2009). Génerallement, un système informatique peut inclure un SBT pour pouvoir effectuer ces enregistrements sous forme d'applications, plate-formes dédiées à cet objectif. Un SBT doit pouvoir enregistrer l'interaction de l'utilisateur avec un système informatique comme sous forme particulière, collecter ces informations enregistrées afin de construire des bases de traces pour visualiser celles-ci si besoin ou bien les exploiter pour produire des connaissances sur le comportement des utilisateurs. Ces bases de traces peuvent faire l'objet de différents types d'analyses que Laflaquière a groupé selon la nature de l'activité étudié (Laflaquière, 2009). Cette activité peut être le système informatique lui même c'est-à-dire l'interface et les fonctionnalités, les comportements et l'interaction des opérateurs dans ce système, ou encore les opérateurs eux-mêmes et leurs activités individuelles ou collectives. Le résultat de cette analyse qui est un modèle de l'activité pour l'évaluation du sujet étudié (Sanderson and Fisher, 1994). Cette base de trace permet la description du domaine étudié en s'appuyant sur l'observation des activités tant dis que l'évaluation de l'activité repose sur l'utilisation d'un ensemble de critères et mesures pour comparer avec le modèle établi. par exemple, les techniques des traces d'interaction sont souvent quantitatives (statistiques) et considèrent ces traces comme des données élémentaires temporelles, alors que le besoin de techniques d'analyse n'a cessé de s'accroitre.

1.6.1 SBT pour l'analyse des systèmes informatiques

Les premières exploitations des SBT avec des systèmes informatiques tend à analyser ces systèmes par leurs concepteurs et administrateurs. les traces sont souvent enregistrés sous forme de fichiers log, produit automatiquement par les systèmes (Système d'exploitation, SGBD, application, Systèmes embarqués, etc.), qui contiennent des informations sur l'exécution des différent instructions de l'application lors du lancement des applications ou authentifications des utilisateurs, etc. De ce fait, leur analyse est importante pour optimiser les performances du systèmes, concevoir de nouvelle mise a jour, (Charbonnaud C, 2005), à gérer les défaillances du système (Ottogalli and Vincent, 1999) (Charbonnaud C, 2005) (Prié, 2006), etc. En effet, avec l'importance de ces traces pour les concepteurs, plusieurs outils approprié qui facilitent leur exploitation, comme: Intel trace analyzer Collector[3], Oracle trace analyzer[4] , etc., et qui génère des statistiques sur le

[3]https://software.intel.com/en-us/intel-trace-analyzer
[4]http://www.oracle.com/technetwork/products/clustering/overview/tracefileanalyzer-2008420.pdf

fonctionnement d'un système. Au niveau conception et développement, l'analyse de ces traces leur facilite la conception (Lee, 1996) et le développement (Etiévant, 2004), aussi le débogage et le test de l'application (Kamin, 1990) (Bates, 1995) (Ducasse et al., 2006). Cette analyse permet de détecter les erreurs de codage et améliorer certains concepts de l'application.

Avec l'émergence des interfaces graphiques et leurs conceptions, l'analyse des systèmes à base d'interface homme machine (IHM) est devenu de plus en plus répandu. L'analyse de ces systèmes informatiques s'est focalisée sur l'étude de l'ergonomie de celles-ci mais sans ignorer les contenus fonctionnels accessibles par le bias de ces interfaces particulièrement dans le Web. Cette analyse est basée principalement sur deux mesures de qualité: l'utilité et l'utilisabilité. Selon Nielsen, l'utilité d'un système mesure l'aptitude du système a fournir a ces utilisateurs les différentes fonctionnalités demandées.

L'utilisabilité mesure la confortabilité du système lors de son usage par utilisateur (Nielsen, 1994). L'un des inconvénients majeur de ce genre d'analyse est qu'elle est couteuse en matière de tests. Par conséquent, les chercheurs en IHM ont développé des techniques de développements de logiciels afin de pouvoir capturer les événements associés aux périphériques d'entrée avec ceux de l'interfacer utilisateur (boutons, liste, mouvement de souris, etc.) (Hilbert and Redmiles, 2000) (Brinkman et al., 2006). Ces traces collectées permettent par la suite une analyse des opérations effectuées et par conséquent une évaluation des qualités et des défauts de l'interface du système utilisé en se référant à des critères prédéfinis ou à un modèle idéal (e.g.identifier les éléments susceptibles de créer des bugs).

Parmi les premiers SBT on trouve Playback (Neal and Simons, 1984), c'est un SBT contenant un dispositif de traçage des interactions utilisateurs qui permet d'analyser a posteriori et pas-à-pas l'ensemble des instructions exécutées lors d'une utilisation grâce a un système d'annotation.La plateforme MacSHAPA synchronise l'enregistrement vidéo et les mouvements de la souris d'un utilisateur en offrant des outils pour traiter (réécrire, encoder, filtrer, etc.) les traces (Sanderson and Fisher, 1994). Néanmoins, il est possible d'effectuer une analyse statistique des traces par le bias des outils simples proposés (Siochi and Ehrich, 1991), comme l'éditeur de commande, qui permet de générer une interprétation automatique des commandes exécutées et les réponses fournies par le système pour qu'après appliquer un algorithme de détection de répétition de modèle pour détecter les problèmes d'exécution. Un autre outil qui permet la visualisation et le traitement des logs d'interaction nommé Experiscope (Guimbretière et al., 2007).

Un autre système MIKE (Olsen and Halversen, 1988) permet de grouper les événements de l'interface utilisateur puis les projetés avec les composants de l'interface qui les ont déclenchés et aux instructions de l'application exécutées, par l'intermédiaire d'un modélisateur abstrait de l'interface qui décrits explicitement les liens entre les aspects de l'application et les événements de l'interface, d'une manière que l'analyse des données collectées dans les logs afin de générer des rapports décrivant l'activité. Cette outil offre le choix des opérations à représenter par des digrammes de temps.

Le système CHIME (Badre and Santos, 1991) enregistre aussi les événements issus de l'interaction avec l'IHM du système et les transforme en unités sous forme prédéfinis, ces modèles d'unités sont collectés et réutilisés automatiquement durant l'interprétation dans le but de tester automatiquement l'interface du système. En outre, l'analyse des systèmes a fait émergence aux travaux de conception de ces systèmes.

Avec l'émergence du web 2.0 et la grande croissances du développements des systèmes basé-web, l'analyse de ces systèmes basés-web est devenu un axe de recherche d'actualité ou beaucoup de SBT ont été proposé. la particularité présente dans l'accès a ce type de système est caractérisé par :

- Le système a analyser contient, à la fois, un navigateur et un contenu informationnel
- L'objectif de l'utilisation du système est avant tout l'information alors que pour les systèmes traditionnels c'est la manipulation et la modification de données
- Les utilisateurs de ces systèmes forment une population non spécifique et leurs activités sont mal définies (il ne s'agit plus d'une situation classique de travail).

Donc, l'analyse de ces systèmes ne prends pas en compte L'IHM du système mais Les données enregistré dans les traces qui peuvent être de natures différentes (e.g., logs client/serveur) relatives aux interactions des utilisateurs avec les objets de l'environnement Web (URL, pages, métadonnées, etc.). Ces traces sont exploitées directement par des spécialistes (Dubois et al., 2000), ou via des systèmes d'analyse automatique (Pirolli et al., 2002). Cette analyse permet d'identifier les principaux défaillances et problèmes d'utilisation de ces sites afin de les améliorer pour répondre aux besoins des utilisateurs en prenant en compte les différentes contraintes liés au web.

1.6.2 SBT pour l'analyse des comportements utilisateurs

L'interprétation des traces pour analyser le comportement des utilisateurs est devenu un axe de recherche important avec l'émergence des systèmes distribués. Une multitude de travaux de recherche liés au problème d'analyse de comportement utilisateurs ont été proposés (Srivastava et al., 2000). Une telle analyse dans un tel environnement doit être flexible avec les issus relative au systèmes distribués par exemple l'hétérogénéité des machines. Cependant, un SBT doit inclure un système de monitoring du système distribué qui peut surmonter cette difficulté.

La mise en place d'un système de monitoring dans un SBT est indispensable pour l'analyse des comportements utilisateurs dans un système distribués. Le projet GRUMPs (Gray et al., 2004) illustre un système de monitoring de l'utilisation d'un système distribué en se basant sur les traces d'interactions présentes dans différentes sources(Logs, camera, protocoles thinkalouad). Ce système permet la capture des données et l'analyse de celles-ci en offrant des outils de filtrage, d'abstraction, d'agrégation, de stockage, de visualisation et de générations de rapports

.xml de ces données. Son principe est de fournir un modèle réutilisable ou l'activité relie les données d'interactions aux objectifs d'investigation en cherchant les traces pertinents. Ce systeme a été évalué dans l'analyse les comportements d'un groupe réparti d'étudiants de programmation (Thomas et al., 2003).

Le traçage des navigations est souvent basé sur une recherche d'information ce qui engendre des fichiers logs contenant les traces et qui sont largement utilisés pour l'analyse. Ainsi, ces traces représente le comportement principal à analyser (Hawkey and Inkpen, 2005), ces traces sont souvent appelés « traces de comportement » (Pirolli et al., 2002). Dans ce cas, un SBT permet a la fois d'analyser l'utilisation des interfaces du systèmes et l'analyse de l'usage du contenu informationnel. Avec l'apparition des premiers navigateurs web, de nombreux chercheurs se sont intéressés à l'analyse de l'utilisation en se basant sur les fichiers logs générés par ces derniers (Catledge and Pitkow, 1995) (Tauscher and Greenberg, 1997) (Cockburn et al., 2002).

Weinreich et al. ont étudiés le comportements vingt-cinq internautes observés durant plus de trois mois utilisant le navigateur Firefox (1.0) pour enregistrer les logs de navigation de ses utilisateurs (Weinreich et al., 2006). Ils ont instrumentés cette étude par l'usage des méthodes de traitements systématiques, calculs statistiques, mesures de corrélation entre des variables du log, etc.. Ils ont démontrés grasse a cette analyse qu'il y a une baisse significative de l'utilisation du bouton « retour » par rapport aux études antérieurs, ce qui a été justifié par l'extension de l'utilisation des onglets et de nouvelle instances de fenêtres du navigateur pour maintenir les pages sur lesquelles les internautes pensaient revenir. Une autre exploitation des logs de navigation en se basant sur des mesures statistiques, le nombre de pages visités, le nombre de fenêtres actives, etc., a été mené. Les résultats obtenus ont été graphiquement représentés pour mettre en évidence les concepts temporels susceptibles qui caractérise i,e navigation observé (Hawkey and Inkpen, 2005).

L'objectif de l'analyse du comportement utilisateurs avec le contenu web est pour étudier les parcours des internautes ou le trafic des données sur le web, tant dis que cette analyse peut faire l'objectif comme la prédiction des actions utilisateurs et systèmes de recommandations. Dans ce genre d'études, l'analyse ne se limite pas que sur les traces de navigation du navigateurs autre types d'information est requises comme l'historique de navigation qui peut donner une vision claire sur l'exploitation des ressources web. Ce genre de traces peuvent provenir de nombreuses sources sur le web (serveurs, caches, logiciels serveur, etc.). Le log d'un serveur par exemple conserve les traces des requêtes et opérations exécutés estampillés par les dates et les heures d'exécution (Hanoune and Benabbou, 2006). Ces traces peuvent être exploités et visualisé graphiquement par des outils poussées tels que: WebAlizer[5], Analog[6], ClickTracks[7],Google Analytics [8] etc.

[5]http://www.webalizer.org/
[6]http://www-verimag.imag.fr/DIST-TOOLS/TEMPO/AMT/content.html
[7]http://www.clicktracks.com/products/pro/index.php
[8]http://www.google.com/analytics/

Dans (Cockburn et al., 2002), les auteurs ont étudiés l'usage du web ou ils ont basés sur le nombre de visite d'une page. D'autre analyseur sophistiqué des logs a été développé dans (Hanoune and Benabbou, 2006) dont ils sont intéressés à la découverte des comportements habituels des internautes via les combinaisons des pages les plus populaires, les navigateurs et les systèmes d'exploitation les plus utilisés, etc.. DIfférent types de travaux qui se sont intéressés à des objectifs plus particuliers tels que: l'analyse de l'impact de la musique sur les internautes (Galan, 2002), l'étude du trafic des données sur le Web (Patarin, 1999). La problématique de la sémantique des parcours Web a été étudiée dans (Beauvisage, 2004) en appliquant des traitements statistiques sur les traces afin de reconstruire les parcours via les URLs grâce à une classification similaire a celle des annuaires web.

D'autres travaux (Chi, 2002) (Chen et al., 2004) (Rossi et al., 2005) ont employé des techniques de visualisation des traces de navigation pour identifier les points critiques des sites Web ou les cheminements les plus utilisé pour atteindre une page cible afin de rendre la navigation plus facile aux utilisateurs. Néanmoins, nous ne pouvons pas l'analyse des traces de navigation caractérise le comportement de l'internaute comme ses objectifs et ses taches restent cachés (Weinreich et al., 2006). Par conséquent, cela a imposé les chercheurs à analyser le comportement des utilisateurs en fonction de leurs activités.

1.6.3 SBT pour l'analyse des utilisateurs et de leurs activités

L'analyse des traces des utilisateurs avec des SBT peut servir l'interprétation de ces utilisateurs ou leurs activités. Cette analyse permet la caractérisation des utilisateurs via des modèles de traces durant une interaction utilisateur avec un SBT représentant son activité peut refléter les processus mentaux et comportementaux de cet utilisateur ce qui permet de comprendre l'activité effectuée. En effet, cette analyse requiert une observation de cet utilisateur durant son interaction. La création des modèles plus ou moins formels a base d'interprétation des traces représente généralement une valeur ajouté au système (Halvey et al., 2005). Toutefois, les travaux de modélisation se basent généralement sur l'inférence des taches, par conséquent, différentes techniques sont utilisés comme les systèmes experts, et la fouille de traces par apprentissage automatique.

Les traces pour l'analyse de l'activité une activité, en informatique ou une grande masse de traces est enregistrées, donc la nécessité d'un modèle référence de l'activité. Cependant, il est souvent très difficile de disposer à priori d'un tel modèle,sauf dans des cas exceptionnelle. Pour les activités médicales dans le dossier patient et SIH par exemple, les critères d'évaluation de telles activités peuvent être définis en fonction de diagnostic et la qualité des soins visés par l'opérateur. Cela rend possible la conception de modèle représentant l'activité autour de ces critères. L'objectif est donc de construire un modèle explicatif prédictif pour analyser les traces afin de modéliser l'activité de l'opérateur.

Dans le cas général, l'analyse de l'activité est caractérisé par l'identification des indicateurs caractérisant cette activité. Ces indicateurs sont des variables auxquelles sont attribuées des pro-

priétés (Halvey et al., 2005), ces indicateurs permettent de décrire les traces de l'activité qui sont exploitées pour la construction du modèle. Par exemple, le nombre de patient admis dans un service hospitalier, ou le temps passé dans un service hospitalier forme des indicateurs de traces de l'activité hospitalière relative au dossier patient. Bien que ces indicateurs soient largement utilisés pour caractériser des activités médicales, d'autres indicateurs sont également utilisés pour caractériser d'autres types d'activités telles que celles de recherche d'information sur le Web (Jansen and Pooch, 2001) par conséquent ces indicateurs sont relatives à la nature de l'activité étudiée.

Toutefois, le fait que les indicateurs soient des mesures quantitatives produites par des calculs statistiques sur les traces d'activité, met en doute leur capacité à fournir une analyse qualitative de cette activité.

1.7 Conclusion

Nous avons présenté dans ce chapitre les informations nécessaires pour cerner les concepts relatifs à l'analyse de traces, leurs propriétés ainsi que la récolte et les techniques d'analyse. Nous avons effectué en ce sens une description de la trace en mettant l'accent sur l'activité qui est la source d'une trace, par ailleurs la connaissance intervient comme l'élément qui valorise les traces collectées, d'où l'effort des techniques de fouille de données et l'aide à la décision.

Nous nous intéressons dans le chapitre suivant à l'introduction de l'analyse des traces médicales dans le cadre de cette thèse pour développer un système d'aide à la décision clinique basés traces.

Chapitre 2

Fouille de Données pour l'Analyse des Traces Patients

Sommaire

2.1 Introduction

Avec la complexité scientifique croissante dans ces dernières années, une grande quantité de traces dans le domaine médical est enregistré sous forme électronique sous le nom du dossier patient électronique. Ces traces sont collectées et utilisées pour différentes objectives. Ces traces sont enregistrées et utilisés principalement pour la gestion et l'analyse de la population. Par exemple, l'existence de traces épidémiologiques précises permet le suivi et l'analyse des conditions sanitaires et sociales de la population. Certaines traces couvrent toute la population, et sont recueillis pendant des décennies. Ils sont fréquemment utilisés pour la recherche, l'évaluation, la planification et à d'autres fins par divers utilisateurs en termes d'analyse et de prévision de l'état de santé des individus.

Une trace dans le domaine médical s'avère capitale pour le patient, mais aussi pour le praticien. Il est indispensable que la trace de chacune des observations, de chaque acte puisse être retrouvée. La traçabilité des patients est une obligation qui vise à garantir une meilleure prise en charge du patient et à assurer la continuité des soins. La traçabilité a pour objectif :

- l'identification du patient,
- l'identification des professionnels qui ont promulgué des soins au patient,
- de retrouver toutes les informations du processus de soins promulgués au patient.
- protéger les patients,
- protéger les acteurs de la santé,
- améliorer la qualité des soins,
- améliorer le contrôle des soins,
- faciliter la recherche d'informations.

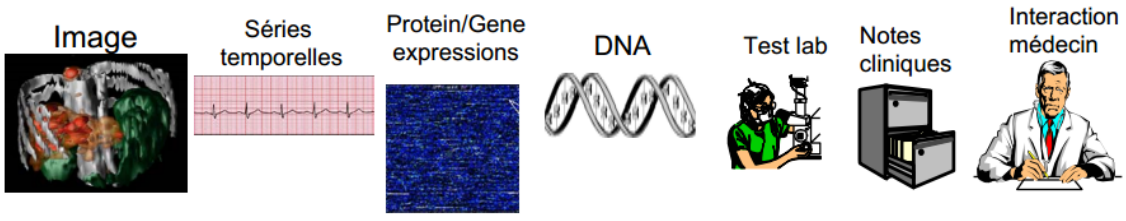


Figure 2.1: Exemples de traces patients

Le dossier patient est le portail de recueil et de conservation des traces administratives, médicales et paramédicales des patients. Le dossier patient assure la traçabilité de toutes actions effectuées. C'est un outil de communication et de coordination entre les praticiens médicaux. Il permet le

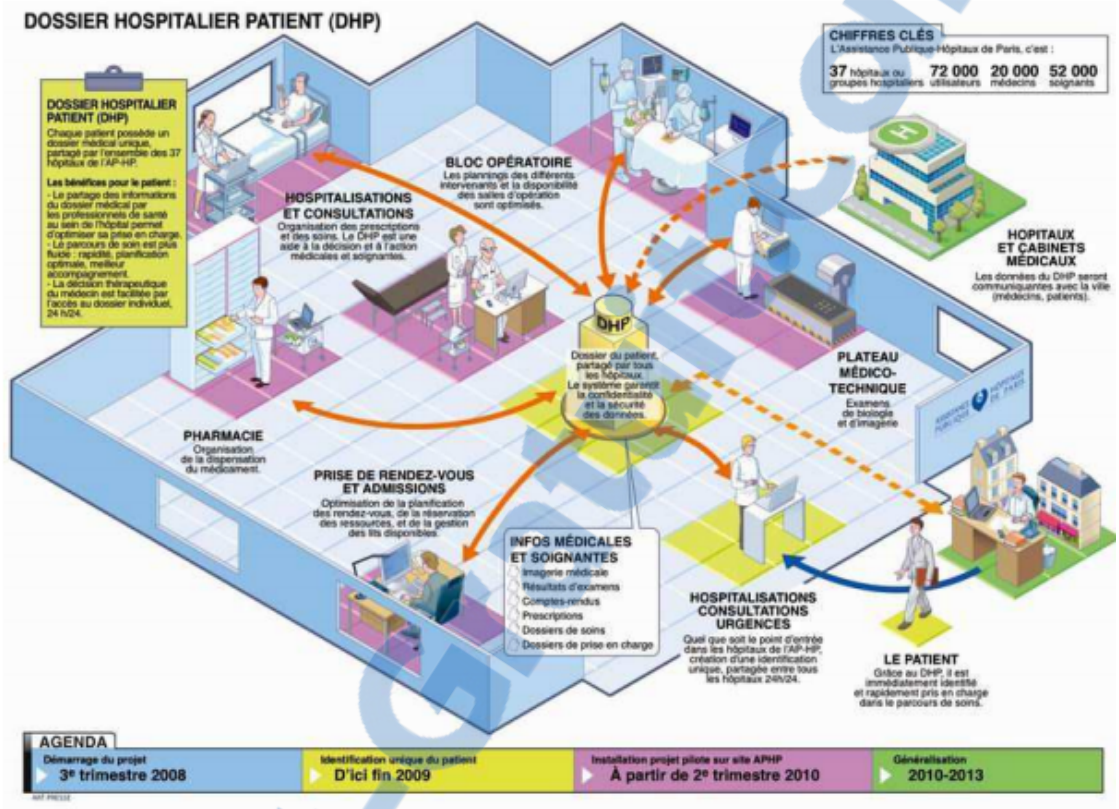


Figure 2.2: Vue d'ensemble du dossier patient(Source site APHP[1])

suivi et la compréhension de l'état de santé du patient. C'est un élément indispensable pour améliorer l'état de santé d'une population (Figure 2.2). L'ensemble de traces patient ou bien le dossier patient gère les processus cliniques et reçoit les informations cliniques par les plateaux techniques tels que les laboratoires, les blocs opératoires, radiologie la pharmacie qui sont attributaires de services pour les unités de soins (Degoulet and Fieschi, 1998). La figure2.2 présente un exemple du rôle du dossier patient dans le système de santé.

La traçabilité dans le dossier patient peut prendre des formes différentes : narratives (ex. description de l'état clinique et/ou d'un comportement), structurée à partir de nomenclatures reconnues (ex. diagnostics infirmiers), graphique (ex. courbe de pression artérielle), mais complémentaires (figure2.1).

L'analyse de trace des patients est un processus important pour le patient et pour la population. Elle sert à limiter les risques et améliore la qualité de la prise en charge du patient. Elle permet d'identifier les anomalies dans le processus de prise en charge du patient afin de les corriger. Elle peut être considérer comme un moyen de prévention pour la population en signalons les différents facteurs de risques pour la population Notamment à partir du contenu des traces (Figure 2.3).

[1] www.aphp.fr/

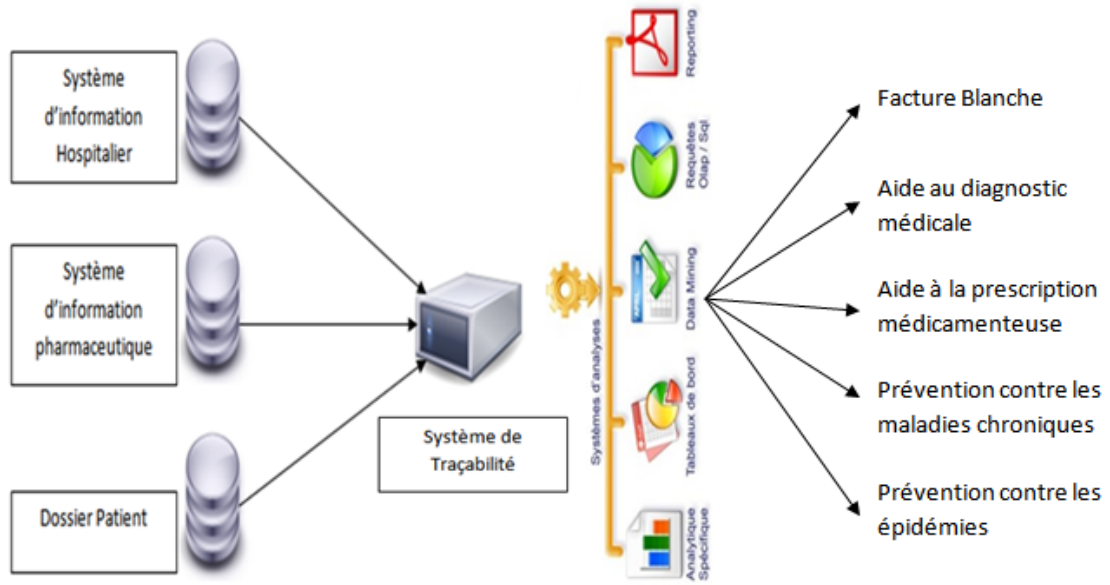


Figure 2.3: Analyseur de traces d'un système de santé

L'utilisation des approches de fouille de traces pour l'analyse des traces et plus particulièrement en traces médicales est devenue inévitable surtout avec l'évolution de ces approches et leurs apports au domaine médicale et spécialement à l'aide de décision médicale. Parmi ces domaines médicaux, nous nous intéressons dans le cadre de ce chapitre aux analyses des traces relatives aux maladies cardiovasculaire CAD qui représentent la première cause de mortalité dans le monde. Par cet article, nous décrivons les caractéristiques d'un système d'aide à la décision clinique CDSS pour le diagnostic du CAD. Nous présentons brièvement les méthodes utilisées pour sa mise en œuvre. L'objectif de ce chapitre est d'expliquer l'apport clinique des CDSSs pour la l'analyse des traces patients.

2.2 Analyse des Traces patients

Aujourd'hui, la quantité des traces des patients collectées dans les bases de traces médicales s'accroit rapidement. L'analyse de ses traces est devenue primordial pour la prise de décision médicale. Il est largement reconnu que l'analyse des traces des patients améliore le système de santé. Deux aspects sont relevés pour motiver l'analyse de ces traces:

1. Les systèmes à base de connaissance peuvent supporter les traces relatives aux patients afin de faciliter la prise de décision

2. La découverte de nouvelle connaissance extraite par l'analyse de ces traces en choisissant un échantillon d'étude représenté par un ensemble de descripteurs (Figure2.1)

A cet effet, avec l'augmentation des tailles des bases de traces qui met en jeu l'insuffisance de l'analyse des traces manuelle, de nouveaux axes de recherches qui tente de rendre cette analyse automatique et intelligente ont étés abordés. Un de ces axes est l'extraction de connaissance à partir des traces ECD (Knowledge discovery in database KDD). L'étape principale dans la découverte de connaissance est celle de la fouille de traces.

Comme beaucoup d'autres taches d'analyses et d'intelligence artificielle, il existe deux approches principales pour l'analyse de traces des patients. La première est l'approche d'ingénierie de connaissances dans laquelle la connaissance de l'expert alimente directement le système sous forme de règles de classification. Cette méthode soulève deux problèmes, l'importance du système cognitif et le gros nécessaire d'investissement (temps et expertises).

L'autre approche est la fouille de traces. La fouille de traces offre un ensemble de méthode d'induction en construisant un classificateur par apprentissage à partir d'un ensemble déjà étiquetés. La plus part des travaux d'analyse et de prédiction utilisent les approches de fouille de traces, qui requiert un échantillon représentatives de la population déjà étiquetés donc elles sont moins couteuses ce qui rend ces méthodes plus adéquates aux problèmes d'analyses (Figure 2.3).

Dans cette thèse nous nous intéressons à la deuxième famille d'approches. Nous focalisons particulièrement aux algorithmes de classification pour la construction des systèmes d'analyse de traces patients. Dans les sections qui suivent nous allons présenter les différentes caractéristiques d'un système d'analyses de traces.

- Collecte et représentation numérique des traces
- Construction d'un système d'analyses de traces
- Évaluation et interprétation d'un système d'analyse de traces.

2.2.1 Collecte et représentation numérique des traces

Afin d'appliquer les techniques de fouille de traces il est nécessaire de recueillir les traces des différentes sources (Base de traces, fichiers, etc.). En effet, pour avoir de bons résultats, il est vital de sélectionner les traces les plus pertinentes qui peuvent représenter la globalité des traces. Ceci nécessite du temps et de ressource humaine considérable, par exemple dans le domaine médical, on a besoin des experts du domaine tels que les médecins, infirmiers et le personnel administratif pour qu'ils se coopèrent avec les analystes afin de construire un Système d'analyse de traces à base de traces pertinentes.

Les traces des patients sont collectées à partir des informations cliniques ou administratif relatifs aux patients, ce qui implique des règles éthique et légale désigné pour protéger la confidentialité de ces traces. On peut définir trois règles à prendre en compte (Camhi, 2004):

- Droit des patients: chaque patient a le droit d'accéder à ses traces cliniques en lui offrant une meilleur qualité de service c'est-à-dire il faut soigneusement tenir ses traces cliniques à jour.
- Confidentialité et sécurité: les traces des patients ne sont accessibles qu'au personnel responsable au patient
- Bénéfices attendues: est-ce que le bénéfice est bien important pour divulguer l'information.

Les traces des patients ne peuvent pas être traitées directement dans le système d'analyse de traces spécialement avec la nature homogène des traces (Figure2.1). Ces derniers ne sont pas capables de les traitées directement, cependant, une étape de représentation numérique est nécessaire.

Le formalisme mathématique le plus répandue est l'utilisation d'un espace vectoriel de comme espace de représentation cible. Ce qui est particulier dans cette représentation est que chaque unité (examens laboratoires, examens cliniques, ADN, etc.) est associée à une seule ou plusieurs dimensions.

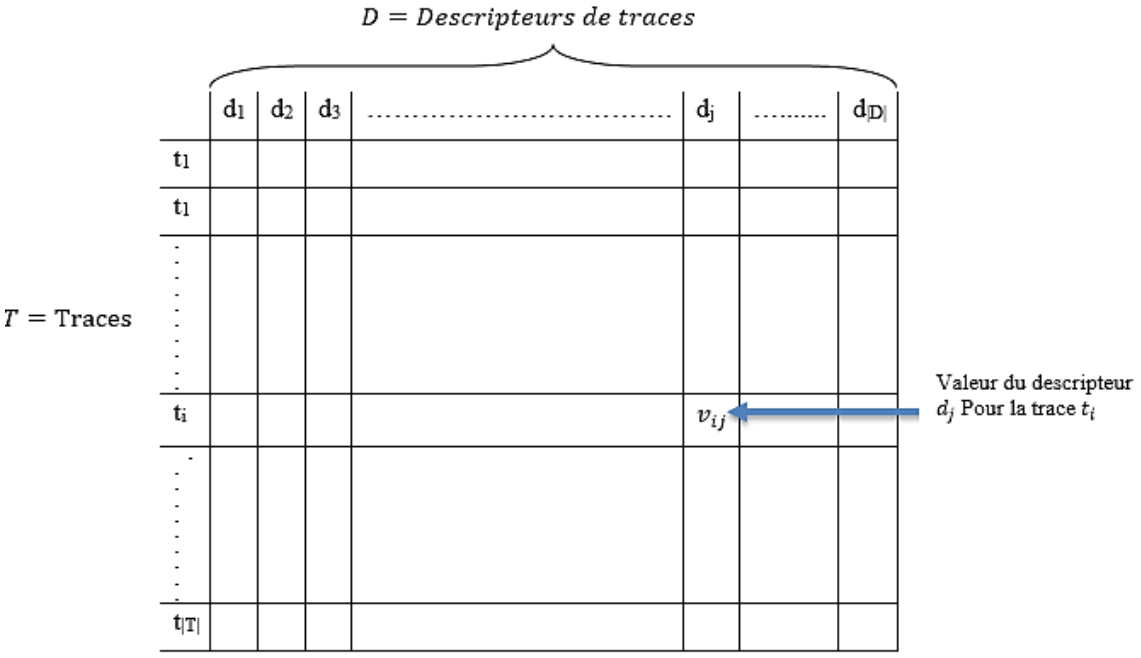


Figure 2.4: Représentation vectorielle des traces

On transforme une trace t_i en un vecteur $\overrightarrow{t_i} = \{v_{i1}, v_{i2}, v_{i3}, \ldots, v_{i|D|}\}$ ou D est un ensemble de descripteurs sélectionnés depuis les différents actes et soins médicaux du patient pendant l'observation de l'activité médicale. La valeur v_{ij} correspond à la valeur d'un acte d_j pour la trace t_i du patient. Cette notation peut couvrir toutes les informations concernant les traces

des patients, par conséquent, c'est la meilleure pour faire l'analyse car elle est adéquate avec les méthodes de fouille de traces voir Figure2.4.

La vectorisation d'un ensemble de trace crée une matrice Traces × Descripteurs ou chaque cellule représente la valeur d'un acte ou bilan dans une trace. Cette matrice peut être volumineuse c'est pourquoi un mécanise de réduction pour en extraire les valeurs pertinentes s'avère indispensable afin de réduire la complexité du système d'analyse de trace.

2.3 Construction du système d'analyse de trace

2.3.1 Prétraitement des traces

Avant d'appliquer les techniques de fouille de traces, le prétraitement des traces est indispensable, ce processus inclus plusieurs étapes pour assurer une bonne qualité des traces et dégager les erreurs relatives, la redondance et prendre en compte la confidentialité des traces des patients (par exemple retirer les noms et les identifiants des patients). Les patients sont concernés pour la préservation de leurs traces personnelles. La fouille de ces traces doit gérer les traces d'une façon anonyme avant de commencer l'analyse. (Nakao et al., 2005).

Les traces des patients ont plusieurs dimension alors certains descripteurs sont plus pertinent que les autres par conséquent le choix ces descripteurs, les plus pertinents, est très important non seulement pour réduire le cout du traitement mais aussi pour améliorer le modèle construit (voir chapitre Analyse de la pertinence des traces et réduction de dimension pour la classification).

2.3.2 Fouille de données

Ces dernières années, les technologies de l'informatique sont devenues de plus en plus répandue dans la santé pour répondre aux besoins des cliniciens et pour offrir des outils de support dans la prise de décision clinique. À cette révérence, les outils de fouille de traces peuvent être utiles pour contrôler les limites humaines telles que la subjectivité ou les erreurs dues à la fatigue et à donner des indications prêtes pour les processus de décision (diagnostic précoce et le pronostic, etc.). Les méthodes de fouille de données utilisent des outils informatiques puissants est de grandes bases de traces cliniques, parfois sous la forme de dépôts de traces et entrepôts de traces, pour détecter les tendances dans les traces. Dans les méthodes de fouille de données, on peut choisir parmi une vaste gamme de techniques qui comprennent, entre autres, la classification, clustering, et les règles d'association.

La classification affecte les traces dans des groupes prédéfinis ou classes. Il est souvent désigné comme apprentissage supervisé car les classes sont déterminées avant d'examiner les traces. Les algorithmes de classification exigent que les classes soient définies sur la base de traces. L'une des

applications de la classification dans la santé est le classement automatique des images médicales qui signifie la sélection de la classe appropriée pour une image donnée sur un ensemble de classe prédéfinies.

Ce modèle de classification sera généré automatiquement à partir d'un ensemble traces. Un exemple consiste en la description d'un cas avec la classification correspondante. Par exemple, on dispose d'un ensemble de traces des patients. Un système d'apprentissage doit alors, à partir de cet ensemble de traces, extraire le modèle de classification qui, au vu de la description clinique d'un nouveau patient, devra prédire le diagnostic médical. Il s'agit donc d'induire un classificateur général à partir d'un échantillon de traces. Le problème est donc un problème inductif, il s'agit d'extraire une règle générale à partir de traces observées. Le modèle généré devra classifier correctement les traces de l'échantillon mais surtout avoir un bon pouvoir prédictif pour classifier correctement de nouvelles traces.

La recherche en apprentissage automatique a produit une large gamme d'algorithmes supervisés pour construire des classificateurs (Niharika et al., 2012). Avant de démontrer la performance de notre algorithme on va introduire quelques méthodes de fouille de traces. Nous allons commenter brièvement quelques-unes. Parmi les algorithmes d'apprentissage supervisé existants, on peut faire des regroupements et distinguer de grandes familles : probabilistes, linéaires, neuronales ...

2.3.2.1 Algorithmes probabilistes

Une des approches probabilistes est l'approche bayésienne NB. Le principe de la méthode NB est basé sur le modèle d'indépendance conditionnelle de chaque prédicteur compte tenu de la classe cible. Il attribue une trace à la classe qui a la plus grande probabilité a posteriori en utilisant le théorème de Bayes (2.1). La principale hypothèse de ce genre de méthodes est l'indépendance de fonctionnalités. Ainsi, lorsque les caractéristiques dépendent de l'autre, cet algorithme produit une faible précision de la classification (Lewis, 1998).

$$P(A/B) = \frac{P(B/A) \times P(A)}{P(B)} \tag{2.1}$$

2.3.2.2 Les arbres de décision

Les arbres de décision sont des outils puissants et populaires pour la classification et la prédiction. L'algorithme C4.5 est l'un des importants algorithmes d'arbres de décision Quinlan (1996). L'attractivité des méthodes de C4.5 est notamment en raison de sa capacité à faire face à toutes sortes de données stockées dans les grandes bases de données, dissemblable aux réseaux neuronaux, C4.5 peut être représenté par des règles de types Si-Alors. Il utilise des techniques d'élagage pour optimiser la précision et le ratio de gain pour la sélection des attributs. L'un

de ses paramètres est M, qui est le nombre minimum d'instances qui devrait avoir une feuille. Le second est C, est seuil de confiance, et il est utilisé dans la fonction d'élagage.

2.3.2.3 Algorithmes Linéaires

Ces algorithmes se basent sur un profil (Schölkopf et al., 2001), (Yang and Liu, 1999). SVM méthode est une méthode ML supervisé, utilisé pour la classification qui est largement utilisé pour produire un modèle de prédiction. SVM est basée sur la notion de plans de décision qui définissent les marges de décision. Un Plan de décision est celui qui sépare entre un ensemble d'objets ayant des appartenances de classe. Pour chaque entrée de test donné, SVM prédit laquelle des deux classes possibles constitue l'entrée, dans le sens où elles tentent de séparer l'espace en deux mais certaines manipulations mathématiques les rendent adaptables à des problèmes non linéaires (Platt et al., 1999). Étant donné un ensemble de traces $(x_i, y_i), i = 1, \ldots, r$ ou $x_i \in R^n$ et $y \in [-1, 1]^r$, SVM implique la résolution du problème donné en 2.2

$$min_{w,b,e} = \frac{1}{2} W^T W + C \sum_{i=1}^{r} \xi_i \tag{2.2}$$

avec

$$y_i(W^T \phi(x_i) + b) \geqslant 1 - \xi_i, \xi_i \geqslant 0$$

En (2.2), x_i est un vecteur d'apprentissage et il est mappé dans un espace de dimension plus élevée par la fonction ϕ. Le C est le paramètre de décision pour le terme d'erreur. SVM trouve une séparation linéaire séparant un hyper-plan avec la marge la plus élevée dans l'espace tridimensionnel. En conséquence, la solution de 2.2 ne permet qu'une séparation linéaire. Dans l'opposition, l'utilisation d'une fonction noyau permet une séparation non linéaire (linéaire, polynomiale, base radiale et le noyau sigmoïde). Dans notre thèse, nous utilisons le noyau polynomial.

2.3.2.4 Les réseaux de neurones

MLP est un réseau neuronal feed-forward formé avec le standard algorithme de retropropagation du gradient (Haykin, 1999). MLP est un réseau supervisé, basé sur les données d'entrée, il apprend à partir d'une base de traces afin de construire un modèle de prédiction. En règle générale, l'architecture des MLP se compose d'une couche d'entrée et une couche de sortie avec une ou plusieurs couches cachées. Pour le nœud de chaque couche est liée à chaque nœud dans la couche suivante avec un poids de $x_{i,j}$. Sur la base de l'algorithme de rétro-propagation, la tâche d'apprentissage se produit pendant que les coefficients de pondération de nœuds sont mis à jour et l'erreur de la sortie est comparée au résultat souhaité. Plus explicitement, (2.3) représente l'erreur dans le nœud de sortie j du nième point.

$$e_j(n) = d_j(n) - y_j(n) \tag{2.3}$$

Dans l'équation, d est la valeur désirée et y est la valeur prédite produite par le MLP. Formellement, mettre à jour les coefficients de pondération (2.4) afin de minimiser l'erreur de l'ensemble de sortie sur la base de ceux doit être évalué (2.5).

$$\varepsilon(n) = \frac{1}{2}\sum_{j} e_j^2(n) \tag{2.4}$$

$$\Delta_{ij}(n) = -\delta\frac{\vartheta\varepsilon(n)}{\vartheta v_j(n)}y_i(n) \tag{2.5}$$

Où $y_i(n)$ la sortie du neurone précédent et ϑ est la vitesse d'apprentissage. Dans nos expériences, nous utilisons un taux d'apprentissage de 0,3 et le temps d'apprentissage 500 itérations.

2.3.2.5 Algorithmes à base de traces

Algorithme 1-NN est un algorithme de fouille de traces simple, il utilise une mesure de distance simple pour trouver la trace la plus proche de la trace de test donné, et prédit la même classe que cette trace. Si plusieurs traces sont à la même distance (plus petit) à la trace de test, une première trouvée est utilisée (Aha, 1997).

Le clustering est similaire à la classification à l'exception que les groupes ne sont pas prédéfinis, mais plutôt défini par les traces automatiquement. Le clustering est alternativement appelé apprentissage non supervisé ou de segmentation.

2.3.3 Interprétation des résultats

L'interprétation des résultats nécessite l'examinassions de sortie de l'outil de fouille de traces utilisées afin d'examiner si ce qui été découvrir est intéressent ou pas. Si le résultat n'est pas optimal on peut répéter l'étape de fouille de traces en utilisant de nouveaux paramètres. Comme toute les traces des patients sont complexes, afin d'évaluer les performances d'un modèle, il existe des techniques manuelle, dans ce cas-là les résultats qui amènent les modèles ne vont être validés que par les experts. La deuxième catégorie de techniques de validation est la validation automatique. Par exemple, si nous utilisons plusieurs méthodes de modélisation floue pour traiter les traces. Lorsqu'on interprète les résultats, concernant uniquement la valeur de précision ne révèle pas d'autres informations importantes, comme été démontré par (Cios and Moore, 2002). D'où il existe d'autres mesures de performance, y compris le rappelle, la précision, F-mesure (Van Rijsbergen, 1979) ces mesures se basent sur le fait qu'un modèle est efficace s'il permet d'avoir un maximum de traces bien classés. La Précision (P) mesure la qualité de classification qui est représenté par une fraction des traces bien classés sur tous les traces affectés à la classe et le Rappel (R) mesure la largeur de la classification c'est-à-dire le ratio des traces bien classés par rapport à l'ensemble aux traces appartenant réellement à la classe. La F-mesure est la moyenne

harmonique du Rappel et de la Précision. $F_{mesure} = 2[1/R + 1/P]$

Le graphe ROC Receiver Operating Characteristics sont utilisés pour l'évaluation des performances des classificateurs représente le compromis entre les bénéfices (Vrais Positifs) et les pertes (Faux Positifs) ou l'abscisse est le taux de faux positifs et l'ordonnée est le taux de vrais positifs, il permet de visualiser rapidement la puissance d'un classificateur.

2.4 Domaine d'application et travaux similaires

La construction d'un système capable de reproduire les activités de raisonnement de l'être humain représente le rêve des chercheurs travaillant en intelligence artificielle (Osheroff et al., 2007). C'est la raison pour laquelle la conception des systèmes à base de connaissances capables de réaliser des fonctions de raisonnement symbolique constitue actuellement un champ primordial des recherches. De tels systèmes nécessitent en particulier une représentation adéquate des connaissances mises en jeu, ainsi que des mécanismes efficaces d'exploitation de ces connaissances, ou de raisonnement.

Dans le domaine médical, le raisonnement désigne les stratégies utilisées par les médecins dans l'objectif d'établir un diagnostic en s'appuyant sur les traces hétérogènes disponibles extraites des systèmes d'acquisition. A titre d'exemple, les traces des patients peuvent être les résultats d'un examen clinique, une image, un résultat de laboratoire, un signal, une séquence vidéo, etc. Lorsqu'on traite des traces du monde réel, comme les traces des patients, nous ne pouvons pas occulter l'aspect lié à l'imperfection affectant ces traces.

En effet, les traces médicales souffrent, en général, au moins d'un type d'imperfection comme par exemple l'imprécision, l'incertitude, ou encore, les traces manquantes. Pour ces raisons, les deux aspects qui sont l'hétérogénéité et l'imperfection des traces, doivent être pris en considération dans l'élaboration des systèmes destinés à analyser ces traces. Cependant, le choix du domaine médical à analyser est important afin de produire une meilleure aide à la décision. Dans notre thèse on s'intéresse au domaine de la cardiologie. En effet, la section est subdivisée en deux sous-sections, la première partie décrit le domaine d'application, et la deuxième montre les différents travaux similaires.

2.4.1 Maladies cardiovasculaire

Des études Hendryx and Zullig (2009) ont prouvés que les CAD sont la première cause de mortalité dans les pays sous développements comme l'Algérie (Figure2.5). Il est estimé que 45% des décès enregistrés en Algérie sont dus aux CAD[2]. Le risque d'avoir une CAD est liée par différentes

[2]Kheireddine Merrad,2012, les facteurs de risque liés aux maladies cardiovasculaires, journée commémorative du cinquantenaire de la clinique de cardiologie du CHU Mustapha Pacha, alger. http://www.lexpressiondz.com/linformation _en _ continue/150241-45-des-deces-enregistres-sont-dus-aux-maladies-cardio-vasculaires.html

facteurs comme l'environnement, psychologique, génétiques, les variables démographiques et les services de santé. Plusieurs de ces maladies nécessitent un traitement chirurgical, notamment les maladies coronaires, les vasculopathies. La chirurgie cardio-vasculaire s'est considérablement développée au cours de ces dernières décennies, jusqu'à devenir une des chirurgies les plus répandues dans le monde (pontages coronaires, remplacements valvulaires...)(of Cardiology et al., 1991).

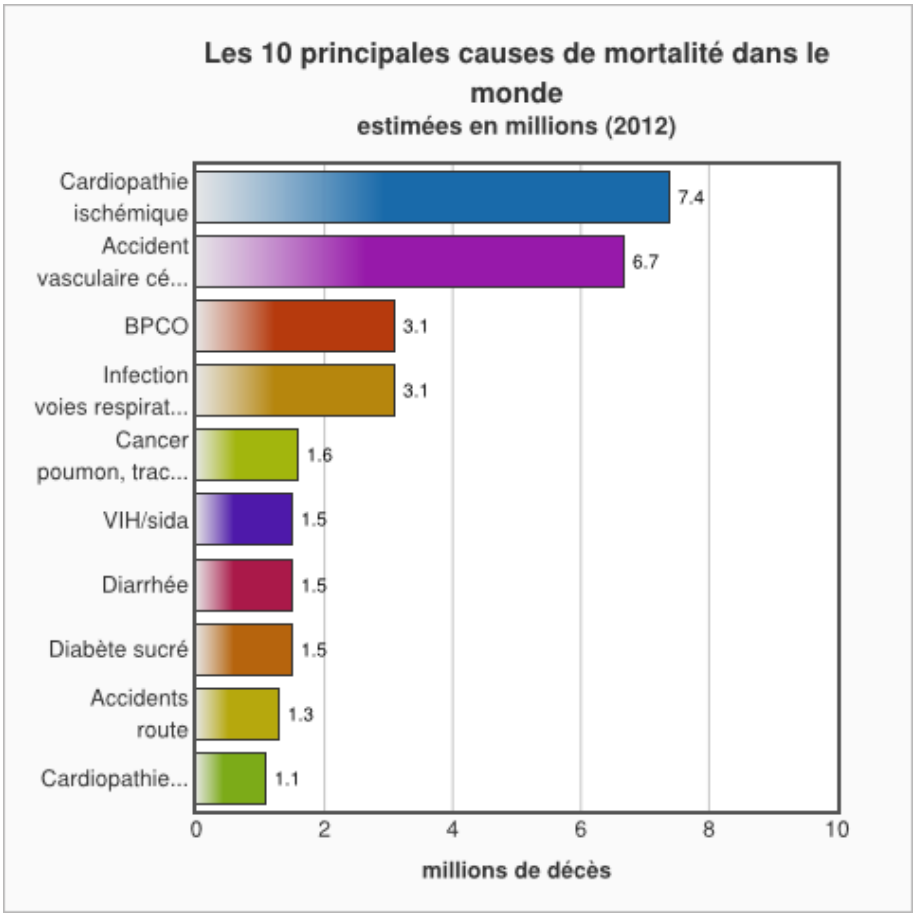


Figure 2.5: Principales causes de mortalités selon l'organisation mondiale de santé

Les procédures chirurgicales conventionnelles exigent une circulation extracorporelle (CEC) et elles s'effectuent par thoracotomie. Mais les grands risques associés à ces procédures (à savoir les séquelles neurologiques, la mortalité, les infections postopératoires) ainsi que leur caractère trop invasif orientent les regards des cardio-chirurgiens vers des alternatives, d'où l'apparition de la chirurgie cardiaque mini-invasive et la chirurgie assistée par robot (Mohr et al., 2001), (Nienaber et al., 1993), (Coste-Manière et al., 2003),(Pike and Gundry, 2003), (Nichol et al., 2008). Ainsi, le pontage coronarien à cœur battant sous endoscopie (Mohr et al., 2001) a entrainé un grand bouleversement dans le domaine de la chirurgie cardiaque, notamment en permettant la prise en charge d'une population de plus en plus âgée. Le grand défi devant les procédures mini-

invasives est la limitation du champ de vue opératoire, ce qui n'est pas le cas dans les méthodes conventionnelles. D'où la nécessité des systèmes de prévention et des alertes pour prédire les facteurs de risques de ces maladies afin de prendre des mesures de prévention. Dans cet article on va présenter les différents Systèmes d'aide à la décision clinique CDSSs, dans la première partie on va décrire le processus de fouille de traces dans le domaine médical, dans la deuxième section on introduit les CAD tout en montrant les techniques médicales usées pour le diagnostic de ces maladies, dans la section qui suit on montre les CDSSs existant pour le diagnostic des CAD.

Le docteur Paul Lépine décrit le processus d'évolution des CAD en trois phases : En premier lieu, l'artère doit subir une inflammation. Puis, dans une tentative maladroite du corps de soigner cette affection, il y aura dépôt de cholestérol et de calcium (durcissement et rétrécissement des artères). Finalement, soit ce dépôt augmentera suffisamment pour boucher l'artère, soit un caillot se formera et obstruera l'artère subitement (thrombose et embolie)[3]

Lorsque les artères coronaires se bouchent ou se durcissent, il se produit une réduction de l'apport en oxygène, nécessaire au cœur. L'athérosclérose, à ne pas confondre avec l'artériosclérose, représente le durcissement et le rétrécissement des grosses et des moyennes artères liées au cœur (aorte, artères coronaires et cérébrales, artères des membres). Elle peut s'accompagner d'un taux de cholestérol élevé et survient lorsque le mécanisme d'élimination des substances graisseuses devient inadéquat. Lorsque cela se produit, la circulation du sang vers certaines parties du corps diminue. La crise cardiaque et l'accident vasculaire cérébral y sont directement attribuables. L'artériosclérose concerne plutôt les petites artères éloignées du cœur, comme les artères musculaires et les artères rénales.

2.4.1.1 Cardiologie

Un cœur normal est une puissante pompe musculaire qui propulse le sang vers les organes, les tissus et les cellules de l'organisme pour apporter l'oxygène et les éléments nutritifs à chaque cellule du corps et retirer le dioxyde de carbone et les déchets produits par ces cellules. Il permet de pomper, au repos, environ 4 à 5 litres de sang par minute.

Le cœur peut être vu comme un moteur musculaire commandé par des impulsions électriques, dont la mission est de pomper une fraction du volume sanguin dans le circuit vasculaire à chaque excitation électrique reçue. Ces impulsions électriques sont sous forme d'une vague de dépolarisation (vague de propagation d'un potentiel d'action) qui débute au niveau du nœud sinusal. Ce dernier joue le rôle d'un «pacemaker naturel». La dépolarisation gagne de proche en proche les deux oreillettes selon une direction générale orientée à gauche. Elle va ensuite traverser le nœud atrioventriculaire avant de gagner le faisceau de His. Après le passage nodo-hissien, débute la dépolarisation ventriculaire au niveau du septum conduite par les deux branches du faisceau de His. Le septum se dépolarise en commençant par son endocarde gauche pour se poursuivre vers son endocarde droit. Enfin, à travers le réseau de Purkinje, l'onde de dépolarisation arrive

[3] http://www.reseauproteus.net/fr

aux deux ventricules en les dépolarisant simultanément (Figure 2.6). Les cellules musculaires (les fibres musculaires) ainsi excitées se contractent dans la direction de leurs fibres, provoquant à chaque battement cardiaque l'éjection de sang des ventricules dans la circulation. Après la contraction, le myocarde doit retrouver ses conditions antérieures au phénomène pour commencer à nouveau le cycle dépolarisation/contraction : c'est la repolarisation.

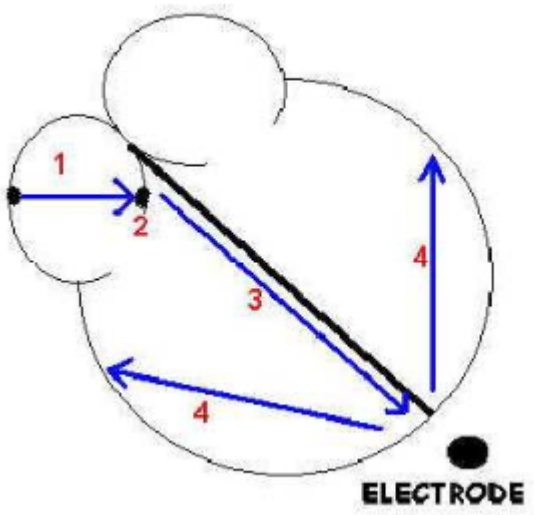


Figure 2.6: Mécanisme de fonctionnement du cœur, (1) Onde de dépolarisation du nœud sinusal au nœud atrioventriculaire, (2) Pause au nœud atrioventriculaire, (3) Onde de dépolarisation du nœud atrioventriculaire à la pointe du septum ventriculaire (conduite par le faisceau de His et ses 2 branches), (4) Onde de dépolarisation de septum à l'ensemble des ventricules (réseau de Purkinje).

Du point de vue fonctionnel, le myocarde ventriculaire est un syncytium c'est-à-dire que les cellules ne sont pas isolées les unes des autres : une excitation qui nait quelque part dans les ventricules conduit, quelle que soit sa localisation, à une contraction complète des deux ventricules. Le couplage excitation-contraction repose sur l'intervention d'une «commande calcique » dans le mécanisme des ponts d'union actine-myosine induisant la contraction musculaire. Cette commande calcique est représentée par la concentration de Ca^{+2} à l'intérieur des cellules musculaires cardiaques, elle-même sous la dépendance directe de la différence de potentiel transmembranaire. A l'échelle macroscopique, l'activité électrique du cœur se mesure de façon non invasive grâce à l'électrocardiogramme (ECG), qui est un tracé de la différence de potentiel électrique entre deux électrodes placées à la surface du corps. Il y a plusieurs dérivations standards, chaque dérivation correspondant à une position de ces deux électrodes de mesure.

2.4.1.2 Troubles cardiovasculaire

« Le cœur est un muscle qui pompe le sang et le dirige vers les organes vitaux. Les battements sont déclenches par des signaux électriques extrêmement précis » (Proulx-Sammut, 1998) Le cote droit du cœur pompe le sang du corps et le transporte, par la voie des artères, vers les poumons. Le sang y trouvera alors de l'oxygène. Ensuite, le côté gauche du cœur reçoit le sang enrichi d'oxygène provenant des poumons et le pompe, à travers l'aorte, dans tout le corps. Les maladies cardiovasculaires (MCV) comprennent une multitude de maladies relatives au cœur et au système circulatoire. Les troubles cardiovasculaires les plus courants sont les troubles coronariens, qui se rapportent aux artères du cœur, et englobent, entre autres, l'angine de poitrine, l'insuffisance cardiaque, l'infarctus du myocarde (crise cardiaque), et les accidents vasculaires cérébraux (AVC) qui se produisent lorsque le cerveau reçoit un apport sanguin inadéquat.

2.4.1.3 Les troubles coronariens

L'athérosclérose, à ne pas confondre avec l'artériosclérose, représente le durcissement et le rétrécissement des grosses et des moyennes artères liées au cœur (aorte, artères coronaires et cérébrales, artères des membres). Elle peut s'accompagner d'un taux de cholestérol élevé et survient lorsque le mécanisme d'élimination des substances graisseuses devient inadéquat. Lorsque cela se produit, la circulation du sang vers certaines parties du corps diminue. La crise cardiaque et l'accident vasculaire cérébral y sont directement attribuables.

L'artériosclérose concerne plutôt les petites artères éloignées du cœur, comme les artères musculaires et les artères rénales.

2.4.1.4 Les accidents vasculaires cérébraux

L'accident vasculaire cérébral (AVC) survient lorsque le cerveau ne reçoit plus un apport sanguin adéquat. Cela entraine la mort de plusieurs cellules nerveuses, ce qui peut causer certaines paralysies. Deux situations principales peuvent déclencher un AVC : l'arrêt de la circulation sanguine par un caillot et une hémorragie cérébrale causée par l'éclatement d'un vaisseau sanguin « malade » dans le cerveau.

2.4.1.5 Facteurs de risque

Plusieurs facteurs peuvent influencer l'apparition de troubles cardiovasculaires. Certains ne sont pas modifiables, comme l'âge et l'hérédité (au moins deux cas de MCV précoces dans la famille), mais d'autres facteurs le sont. Le niveau de risque peut être réduit par la modification du mode de vie. L'état de santé peut également influencer le développement des CADs (voir table 2.1).

Facteurs associes au mode de vie	• Tabagisme • Obésité • Inactivité physique • Consommation excessive d'alcool (deux consommations et plus par jour) • Stress chronique • Prise de contraceptifs oraux (pour les femmes non ménopauses et fumeuses)
Facteurs associes a l'état de santé	• Hypertension artérielle • Fréquence cardiaque élevée • Faible taux de « bon » cholestérol (HDL) • Taux élevé de « mauvais » cholestérol (LDL) • Excès de triglycérides (TG) (graisses de réserve) • Taux élevé d'homocystéine (parfois héréditaire, peut également être du a un apport insuffisant de vitamines B6, B9 et B12 et au stress) • Diabète

Tableau 2.1: Facteurs de Risque du CAD

2.4.1.6 Diagnostique des maladies cardiovasculaire

CADs sont diagnostiqués utilisant un choix d'essais en laboratoire et d'études de représentation. La partie primaire du diagnostic est médicale et des antécédents familiaux du patient, des facteurs de risque, de l'examen matériel et de la coordination de ces découvertes avec les résultats des tests et des procédures. Certains des tests communs employés pour diagnostiquer des CAD comprennent :

EKG/ECG (Électrocardiogramme): C'est un test simple et indolore qui enregistre l'activité électrique du cœur. Le patient est attaché à l'instrument avec plusieurs corrections ou plombs mis au-dessus de sa poitrine, poignets et chevilles. Une petite machine portative enregistre les activités du cœur sur une bande de papier de graphique. Le test affiche comment rapidement le cœur bat et son rythme. La force et la synchronisation des signes électriques pendant qu'elles traversent le cœur sont également vues. Un EKG/ECG peut aider à trouver une crise cardiaque, des crises de l'angine, des arythmies Etc (Blackburn et al., 1960).

Échocardiographie: Ce test utilise le son salue produisent un cinéma du cœur. C'est également un test indolore où une sonde est roulée au-dessus de la poitrine et la machine produit l'image du cœur sur le moniteur. Ceci fournit des informations sur la forme, la taille, les fonctionnements, les soupapes et les cavités du cœur. L'Échocardiographie peut également être combinée avec Doppler pour afficher les zones de l'approvisionnement en sang faible au cœur. Elle affiche les zones du muscle cardiaque qui ne se contractent pas normalement, et des préjudices précédentes au muscle cardiaque (Malergue et al., 1992).

IRM Cardiaque: L'IRM (imagerie par résonance magnétique) Cardiaque ce utilise les ondes radio, les aimants, et un ordinateur pour produire des illustrations du cœur. Ceci donne une image 3D des illustrations mobiles ainsi qu'immobilises du cœur (Crochet et al., 1990).

2.4.2 CDSS pour l'aide au diagnostic des CADs

Les système d'analyse de traces des patients ou bien les système d'aide à la décision clinique (CDSSs) sont «des applications informatiques dont le but est de fournir aux cliniciens en temps et lieux utiles les informations décrivant la situation clinique d'un patient ainsi que les connaissances appropriées à cette situation, correctement filtrées et présentées afin d'améliorer la qualité des soins et la santé des patients » (Abbasi and Kashiyarndi, 2006).

Comme tous les domaines médicaux et afin de prévenir les troubles cardiovasculaires, l'une des solutions possibles est de rendre les gens conscients de leurs risques CAD au préalable et prendre des mesures préventives en conséquence. Selon les experts une détection précoce au stade de l'angine de poitrine peut empêcher un CAD si le médicament approprié est donné par la suite. C'est là que réside l'importance de développer un système pour le CAD. Des études qui ont été faites pour l'étude des facteurs de risques concernant le CAD (Wung et al., 2013)(Vaisi-Raygani et al., 2010), d'autres travaux qui tentent d'analyser les 12-lead ECG (Yang et al., 2000), (Gibler et al., 2005) et 18-lead ECG (Wung et al., 2013).

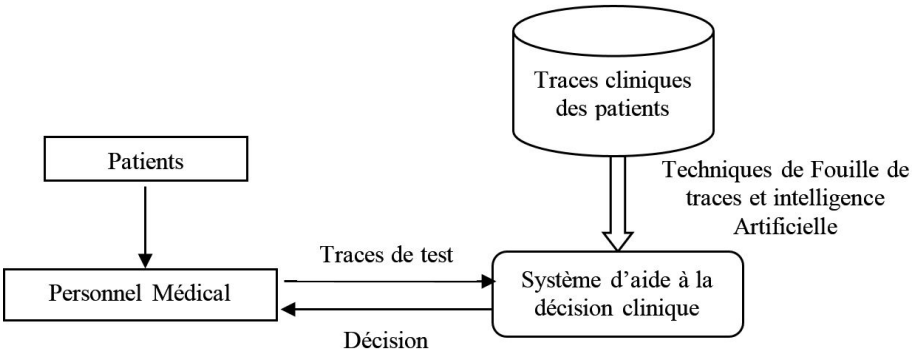


Figure 2.7: Architecture d'un Système d'aide à la décision Clinique CDSS

Les CDSSs concernent les différentes spécialités médicales avec les modalités d'intervention spécifique a chaque spécialités (Dipchand et al., 2007). Par ailleurs, un CDSS doit couvrir certaines activités médicales: prévention, dépistage, diagnostic, Prescriptions etc. Ces activités sont relatives a des spécialités médicales relativement à des maladies chroniques (cancer, diabète, CAD...), les affections aigües et les urgences. Le CDSS doit etre manipulable par différents catégorie de médecins: généralistes, spécialistes, médecins en formation selon les modes d'exercice de soins: cabinets médicaux, services d'hospitalisation, de consultation ou d'urgence des établissements de santé publics ou privés. Leurs modalités d'intervention sont diverses :

- Accès en ligne à des informations de référence dans le contexte d'une situation clinique donnée
- Recherche et présentation des traces cliniques pertinentes dans le contexte de la tâche en cours : décision diagnostique ou thérapeutique, prescription médicamenteuse, tableaux de bords pour le suivi des traitements etc.
- Aide à la documentation des soins sous la forme de listes de traces cliniques pertinentes :
  - Recueillies afin d'établir un diagnostic ou un pronostic ou de suivre les effets d'un traitement,
  - Associées à des contrôles automatiques de la qualité des traces saisies
- Aide à la prescription des actes diagnostiques ou des médicaments au moyen de formulaires établis à partir des recommandations de pratiques et proposant des bilans ou protocoles appropriés à la situation clinique du patient.
- Fonctions de gestion de protocoles pour la prise en charge de maladies chroniques utilisant les diverses modalités d'intervention des CDSS
- Alertes informant les cliniciens de la survenue d'évènements, tels que l'identification d'un résultat d'examen anormal, la détection d'une allergie ou d'une interaction médicamenteuse dangereuse.
- Rappels ou « aide-mémoire» rappelant à l'utilisateur:
  - Soit des recommandations pour la prévention primaire ou secondaire
  - Soit des recommandations pour le diagnostic, la prescription d'examens ou de médicaments, la surveillance d'un traitement

2.4.3 Types des CDSSs

Les CDSSs sont classés en deux catégories : Systèmes à base de connaissances et non basés connaissances (Abbasi and Kashiyarndi, 2006).

2.4.3.1 Systèmes à base de connaissances

Les premiers systèmes d'aide à la décision clinique sont conçus à base des systèmes experts, le but était de formaliser le raisonnement du clinicien sous forme de règles et les intégrer dans un système (DeBusk et al., 1994). Jusqu'à maintenant de nombreux systèmes experts médicaux ont été développés comme MYCIN (Langlotz and Shortliffe, 1983), INTERNIST (van Melle et al., 1984), CADIAG-2 (Iantovics et al.). De premières recherches montrent que ces systèmes pourrait être utile, qu'ils pourraient être utilisés pour assisté les cliniciens dans la prise de décision en

les avertissant des problèmes potentiels, ou de fournir des suggestions pour les cliniciens. (Kerr et al., 2012)

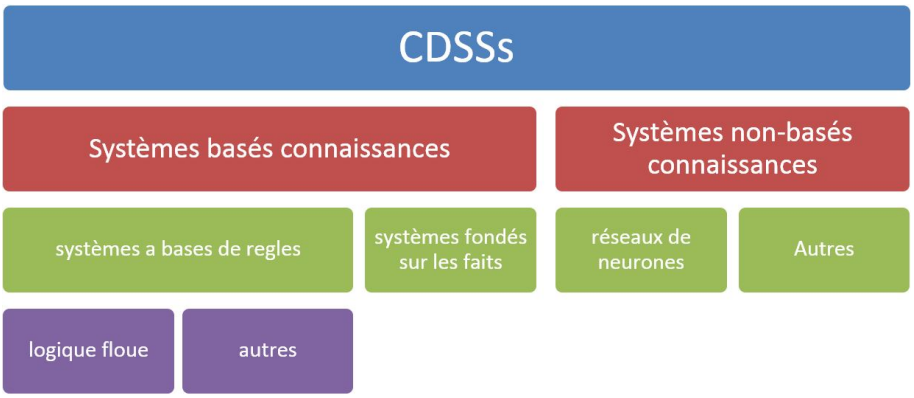


Figure 2.8: types des systèmes d'aide à la décision Clinique

Ces CDSSs sont des systèmes à base de connaissance. D'autres systèmes intègrent des modèles explicatifs qui sont utiles pour l'enseignement dans le domaine médical. Il existe plusieurs descriptions des types des CDSSs. (Anooj, 2012).

Les CDSSs basés connaissance contient souvent les règles sous formes de règles si-alors. Les traces sont généralement associées à ces règles. Les systèmes basée connaissances généralement se compose de trois parties principales. Base de connaissances, règles d'inférence et un mécanisme de communication. Ils sont le type le plus courant du CDSS utilisé dans les cliniques et les hôpitaux. Ils peuvent avoir des connaissances cliniques sur une tâche particulière définie, ou peuvent même être intégrer dans un système base de cas. Nous distinguons deux types de CDSS basée connaissances : les systèmes classiques et les systèmes à base de logique floue.

Les systèmes à base de règles et les systèmes médicaux à base de faits Ces systèmes ont tendance à capturer les connaissances des experts du domaine en expressions qui peuvent être évaluées comme des règles. Quand un grand nombre de règles intégrer dans la base de règles, Il est utile pour stocker les traces pertinentes pour avoir un meilleur modèle pour la prédiction. Les systèmes médicaux à base de faits semblent des outils parfaits en améliorant les soins cliniques ainsi que la bonne prise en charge des patients. Ils améliorent la qualité et la sécurité tout en réduisant le coût (Kaushal et al., 2003).

Un prototype de système intelligent pour la prédiction des CAD a été proposé par Palaniappan and Awang (2008) utilisant les techniques de fouille de traces telles que les réseaux de neurones, les arbres de décisions et l'algorithme de Naïve Bayes. Ce système peut répondre à des requêtes complexes de types « what if » ce que la plupart des CDSSs ne peuvent pas traiter. Il offre la possibilité de prévoir si un patient est suspecté d'avoir une CAD en se basant sur des profils médicaux comme l'âge, sexe, la tension artérielle, le taux de glycémie dans le sang. Le système

est à base de web, facile à manipuler, évolutive, fiable et extensible.

Différentes approches de fouille de traces ont été étudiées, l'arbre de décision et la régression logistique ont étés utilisé pour le diagnostic de l'infarctus du myocarde (Kurt et al., 2008), ils ont utilisés ROC pour évaluer les deux méthodes, d'autres travaux ont étés utilisé pour la comparaison des règles d'association et avec l'arbre de décision (Nahar et al., 2013), les règles d'association et les règles de filtrage (Karaolis et al., 2008). Kochurani et al. (2007) ont proposé une méthode pour le diagnostic CAD basé sur le modèle neuro-flou. (Karoli et al., 2012) et Karaolis et al. (2008) ont développé un système prédire trois pathologies coronaire : infarctus du myocarde, l'intervention coronarienne percutanée et pontage coronarien ils ont utilisé l'algorithme C4.5 pour la classification des différents pathologies. Kurt et al. (2008) ont fait une étude comparative de la régression logistique, la classification et arbre de régression, réseaux de neurones et Carte d'auto-organisation pour développer des modèles afin de prévoir l'absence ou la présence de maladie de l'artère coronaire. Guo et al. (2014) ont proposé un algorithme k-prototype floue et ils ont appliqués leur méthode sur des traces cardiaques.

Systèmes à base de logique floue Les systèmes à base de logique règle floue sont très efficaces en matière de prédiction du diagnostic. La logique floue facilite le traitement de l'imprécision dans le système de prise de décision médicale. Elle peut décrire l'imprécision et l'incertitude dans un langage mathématique précis, afin de représenter explicitement l'imprécision des décisions clinique (Charbonnaud C, 2005).

Multiple travaux ont été proposé dans le cadre des CDSSs à base de logique floue. Dans (Anooj, 2012), l'auteur a proposé un CDSS pour la détection des CAD. Ce système consiste deux phases, la première est celle de la génération des règles floues, et la deuxième est la construction du système d'aide au diagnostic à base de règles générées. La première comprend la génération des règles floues qui vont être pondérer après et intégrer dans le system. Ce système a été testé sur différents data set. Des expérimentations ont été levées afin de montrer l'efficacité du système par rapport à d'autres à base du réseau de neurone par exemple.

Pour le diagnostic de l'artère coronaire Tsipouras et al. (2008) ont proposé un CDSS basé sur des règles floues. Ce système comporte quatre phases : 1) construction d'un arbre de décision a partir des traces, 2) l'extraction des règles à partir de l'arbre, 3) la transformation des règles de la forme normale a la forme floue, 4) et l'optimisation du model. Le système a été testé sur un ensemble de 199 individus décrits par 19 attributs. Ils ont utilisés Cross validation pour valider le système. En calculant la précision et le rappel de la première phase, ils ont eu 54% et 65% respectivement. Lors de la dernière phase ils ont eu une bonne amélioration puisque la précision est de 80% et le rappel et de 65%. Un autre travail dans ce cadre, (Setiawan et al., 2009) ont développés un CDSS flou pour cela ils ont utilisés le data set obtenue de l'université d'Irvine en Californie. Ils ont utilisé une méthode d'extraction de règles sur un ensemble d'approximation. Après ils ont discrétisés les valeurs numériques, ils ont après transformés les règles sous la forme floues. Dans la phase finale

ils ont pondérés les règles. Le système a été testé sur différents ensembles de traces. Les résultats ont montrés que le système est capable de prédire le pourcentage l'artère coronaire mieux que les cardiologues et l'angiographie. Le système est validé par trois experts cardiologues.

Un Autre système qui été développée dans (Bhatla and Jyoti, 2012) en utilisant les techniques fouille de traces et la logique floue sur maltab, ils ont utilisés 4 biomarkers pour la prédiction, et ils ont générer les règles utilisant les deux algorithmes Naïve bayes et l'algorithme d'arbre de décision C4.5 et par la suite ils ont appliqué la Fuzzification des règles. Ils ont montrés que l'utilisation de ces quatre attributs peux augmenter la précision du système jusqu'au 100%.

Dans le travail de Khatibi and Montazer (2010), un moteur d'inférence floue nommé Fuzzy-Inference Hybrid Engine a été proposé utilisant la théorie de Dempster-Shafer (Shafer et al., 1976) et (Dempster, 1967) est la théorie des ensembles flous. Ce moteur hybride fonctionne en deux phases. Dans la première phase, ils modélisent l'imprécision des informations d'entrée par le biais de sous-ensembles flous. Dans la suite, ils extraient les règles floues pour le problème, il applique les règles d'inférence floue sur l'ensemble flou acquis pour produire les résultats de la première phase. À la deuxième phase, les résultats obtenus de l'étape précédente sont considérées comme des croyances de base. Ensuite, ils ont calculés les fonctions de croyance et de vraisemblance pour les croyances de base, ils ont obtenus l'intervalle de croyance qui est le résultat final. La collecte d'informations provenant de différentes sources fournissent diverses croyances de base qui devraient être fusionnés pour produire un résultat. Après avoir appliqué le moteur proposé sur la CAD (CAD), il a donné des taux de 91,58% d'exactitude pour sa prédiction correcte.

2.4.3.2 Les systèmes d'aide à la décision non-basés connaissances

Les CDSSs non-basés sur les connaissances sont basés sur les techniques d'intelligence artificielle. Ces systèmes sont généralement divisés en deux catégories principales: les réseaux de neurones et les systèmes basés sur d'autres techniques de fouille de traces.

Pour trouver des corrélations entre les symptômes et le diagnostic, les réseaux de neurones utilisent les nœuds et des connexions pondérées entre eux, ce qui n'exige pas à la création des règles au préalable. Le processus d'apprentissage du réseau de neurone dans lequel on doit regrouper les individus selon des catégories utilisant des informations pertinentes correspondant à la classe (Kim et al., 1999). Les réseaux de neurones ont été largement appliqués au problème statistique non-linéaire de et appliqués sur les bases de traces médicales. L'objectif de l'apprentissage est d'optimiser les performances du réseau pour avoir une meilleur prédiction du diagnostic pour un ensemble de traces d'entrées.

Ces systèmes boite noir, les réseaux de neurone, Patil and Kumaraswamy (2009) ont proposés un système intelligent pour la prédiction des crises cardiaques, afin de rendre les traces prêtes à être analysé ils ont les regrouper dans un entrepôt de traces. Une fois les traces sont dans

l'entrepôt de traces, ils ont construit des clusters utilisant la méthode K-means afin de construire des groupements des individus similaires. Par conséquent, avec l'aide de l'algorithme MAFIA, les séquences fréquentes appropriés pour la prédication des crises cardiaques ont été extraites. Avec l'usage des séquences fréquentes comme ensemble d'apprentissage et l'algorithme de back propagation, le réseau de neurone a été utilisé. Les résultats étaient satisfaisants en matière de prédiction.

Fidele et al. (2009) a fait appel aux techniques de l'intelligence artificielle comme étant la base du système d'évaluations des facteurs de risques pour les CAD. Le réseau de neurone artificiel comprend un perceptron de deux couches qui emploie l'algorithme de Levenberg–Marquardt et l'algorithme de rétro propagation. Ils ont montrés la qualité de leur système en l'appliquant sur l'ensemble de traces Long Beach.

L'efficacité des réseaux de neurones flous pour prédire l'artère coronaire a été évaluée dans (Abidin et al., 2009) en se basant sur des marqueurs biologiques, leurs habitudes et des profils démographiques. Les performances de prévision des modèles de réseaux neuronaux flous ont été calculées en termes de précision par rapport à la performance de prédiction des modèles de régression logistique. Les résultats ont illustrés pour la prédiction des maladies de l'artère coronaire, il faut prendre quatre marqueurs en considération : indice de masse corporelle, la pression artérielle systolique, le taux de cholestérol et l'âge.

D'autres systèmes non basés connaissances dans (TREE, 2014), une étude a été faite pour la prédiction des maladies de l'artère coronaire en se basant sur les traces collectées d'un centre de recherche, ces traces incluent un ensemble de 303 patients. Chaque patient est décrit par un ensemble de paramètre concernant son ECG : onde Q, l'intervalle ST etc. Comme outils de classification ils ont menés leurs expérimentations avec l'algorithme Naïve Bayes, SVM et un algorithme « Ensemble » qui essaye d'hybridé les résultats obtenus des autres algorithmes. Ils ont trouvés que l'algorithme « Ensemble » avait une précision de 88.5%.

Dans cet étude (Jilani et al., 2009), les auteurs ont utilisés des techniques de fouille de traces pour étudier les facteurs qui contribuent de manière significative à la prédiction du risque de syndromes de l'artère coronaire. Ils ont supposés que la classe est le diagnostic - avec des valeurs dichotomiques indiquant la présence ou l'absence de maladie. Ils ont appliqués la régression binaire. L'ensemble des traces a été prise à partir de deux hôpitaux cardiaques de Karachi, au Pakistan. Pour une meilleure performance du modèle, les techniques de réduction des traces comme l'analyse en composantes principales a été appliquée.

2.5 Conclusion

Il est reconnu que le faite qu'un CDSS, lorsqu'il est bien conçu et mis en œuvre, offre un grand potentiel pour améliorer la qualité des soins de santé et peut-être même permet d'augmenter l'efficacité et de réduire les coûts des soins de santé. Un CDSS ne doit pas être considéré comme

un outil pour remplacer un médecin, mais un outil de support complexe qui nécessite un examen attentif de ses objectifs avant de le concevoir.

Comme tous les domaines en médecine, la cardiologie et les CAD s'avèrent les plus importantes puisqu'elles sont la première cause de mortalité au monde. Beaucoup de travaux ont été proposés pour la prévention contre les risques de ces maladies. Ces travaux ont été sous formes des CDSSs. Parmi il y en a qui sont basés sur des base de connaissances qui sont extraites de la population étudiés. D'autres qui proposent des systèmes boite noirs. Et d'autres qui comparent l'efficacité des systèmes.

Dans le cadre des CDSSs pour le diagnostic du CAD, on va proposer une approche basée logique floue ce qui facilite l'interprétation des règles par le personnel médical. L'importance de la découverte des règles floues significatives et pertinentes sans l'aide des experts ouvre la possibilité de révéler de nouvelles connaissances. Les principaux avantages de cette approche, comme un outil d'acquisition de connaissances sont les suivants: (1) un nombre réduit de règles sont obtenus avec une meilleur précision par rapport aux autres travaux (2) les règles obtenues peuvent être facilement interprétés.

Chapitre 3

Analyse de la Pertinence des Variables pour la Réduction de Dimension pour la classification

Sommaire

3.1 Introduction

Les traces utilisées pour concevoir les CDSSs (Bilans, marqueurs biologiques, signaux...) sont souvent caractérisées par un grand nombre d'attributs qui peuvent surpasser le nombre de données. Ce problème connu sous le nom de "la malédiction de la dimension" constitue un défi pour les différents algorithmes de fouille de données. Dans notre approche on va présenter le rôle de la réduction d'attributs dans le processus d'analyse de traces afin d'éliminer la redondance et les corrélations entre ces descripteurs. Considérer un nombre élevé d'attributs d'une part augmente le risque de prendre en considération des attributs redondants ou corrélés ce qui rend ces algorithmes plus complexes et parfois moins performants. Il est alors nécessaire de procéder à une étape de réduction de la dimension de l'espace des attributs d'entrée. Ce chapitre introduit le problème de sélection d'attributs et les différentes techniques existantes dans la littérature.

3.2 Sélection d'attributs

La fouille de données a était largement utilisée pour l'Analyse des traces et pour le développement des CDSSs. Les réseaux de neurones (Oreski and Oreski, 2014), machines a support vectoriel (Nalinipriya et al., 2012), les algorithmes génétiques (GA), logique floue (Hong et al., 2014) et d'autres systèmes hybrides (Kabir et al., 2011) sont les principaux axes de recherche pour développer ces systèmes. Généralement un processus de développement d'un CDSS consiste deux étapes : le prétraitement des traces et la construction du modèle de trace. Usuellement, le prétraitement implique la collection, la définition des descripteurs et la préparation de traces appropriées tandis que le model de classification est utilisé pour discernement des traces selon différentes étiquettes de classes. Le but de l'étape de prétraitement est et de sélectionner les meilleurs traces à partir d'une base de données et les transformés dans un format appropriés selon la méthode de classification choisit. Cette étape prétraitement permet de rendre l'ensemble des traces plus représentatif, de réduire l'espace de stockage nécessaire de ces données, ainsi que le temps d'apprentissage et d'exploitation des algorithmes de classification.

Les méthodes de sélection d'attributs sont les plus utilisés dans l'étape de prétraitement. La réduction de la dimensionnalité des traces (ou bien sélection d'attributs) permet de rendre l'ensemble des traces plus représentatifs, de réduire l'espace de stockage nécessaire de ces données, ainsi que le temps d'apprentissage et d'exploitation des algorithmes de classification. Les principaux objectifs des méthodes de réduction de dimension peuvent être décrit par:

- Identification des attributs pertinents
- Amélioration de la qualité de la méthode de classification
- Réduction de l'espace de stockage et du temps d'apprentissage

Les méthodes de réduction de la dimension peuvent être divisées en deux grandes catégories : l'extraction d'attributs et la sélection d'attributs (Guyon et al., 2007), (Yu and Liu, 2003) : les méthodes d'extraction d'attributs consistent à transformer l'ensemble d'attributs de départ en un nouvel ensemble d'attributs, généralement plus réduit tout en gardant le plus possible la quantité informationnels des données. Tandis que les méthodes de sélection d'attributs permettent de choisir un sous-ensemble réduit d'attributs à partir de l'ensemble d'attributs originaux (Yu and Liu, 2003).

3.2.1 Extraction de variables

Les méthodes d'extraction de variables à partir d'un ensemble original d'attributs N construisent un ensemble e d'attributs, tels que $e \leqslant N$. Plusieurs méthodes d'extraction de variable existent dans la littérature (Bekkerman et al., 2003). On peut distinguer les méthodes linéaires non-supervisées comme l'Analyse en Composantes Principales (ACP) (Bouroche and Saporta, 1992), les méthodes linéaires supervisées comme l'Analyse Factorielle Discriminante (AFD) (Saporta, 2011), les méthodes non linéaires non supervisées comme l'ACP à noyau (Schölkopf et al., 1998), Locally Linear Embedding (LLE) (Roweis and Saul, 2000), Isometric FeatureMapping (Isomap) et les méthodes non linéaires supervisées comme l'Analyse Factorielle Discriminante à noyau, etc.

On considère un espace d'observations θ dans R^n et l'espace de caractéristiques M relié par l'espace d'observation par une application ϑ :

$$\omega : \theta \rightarrow M$$

$$x \rightarrow \vartheta(x)$$

L'ensemble d'attributs est représenter par un ensemble finit de points x_i. Dans le cadre supervisé, cet ensemble sera modélisé par des couples (point, label) (x_i, y_i) (Voir Figure3.1). Les méthodes d'extraction d'attributs peuvent être distinguées en :

- Méthodes linéaires non-supervisées, la plus célèbre méthode de cette famille est l'Analyse en Composantes Principales (ACP) (Hotelling, 1933; Loève and Lévy, 1948). L'ACP est la transformation linéaire optimale elle permet de garder le sous espace qui a la plus grande variance. Cependant, elle nécessite beaucoup de calculs. La réduction de dimension implique une perte d'informations d'où l'objectif est de trouver le sous-espace qui permet de perdre le moins d'informations.

- Méthodes linéaires supervisées, comme l'Analyse Factorielle Discriminante (AFD). Bien qu'étant une analyse factorielle supervisée, l'AFD est une méthode descriptive et prédictive fondée sur un modèle paramétrique qui discrimine les individus selon des classes connus. Le but de la méthode est alors, comme en ACP, de réduire la dimension de l'espace d'attributs en cherchant de nouveaux attributs, combinaisons linéaires des attributs initi-

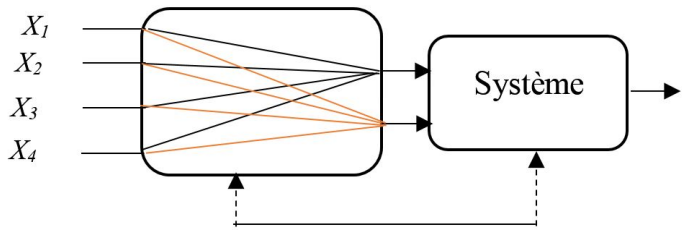


Figure 3.1: Extraction d'Attributs (Guérif, 2006)

aux, suivant lesquels les nuages de points correspondant aux différentes classes sont les mieux séparés et les plus compacts. Les axes factoriels discriminants successifs sont alors déterminés (Bouroche and Saporta, 1992).

- Les méthodes non linéaires non supervisées, parmi ces méthodes on peut citer la fameuse méthode : l'ACP à noyau en anglais Kernal Principal Component Analysis (KPCA) (Schölkopf et al., 1998) qui est une première extension de la méthode ACP. L'ACP à noyau effectue une ACP dans l'espace K appelé espace de caractéristiques à l'aide d'une fonction noyau. qui utilise un noyau pour représenter les données dans un espace de grande dimension (éventuellement infini) où on effectue une projection par l'ACP classique. D'autres méthodes existent comme Locally Linear Embedding (LLE) (Roweis and Saul, 2000), Isometric FeatureMapping (Isomap).

- Les méthodes non linéaires supervisées, comme l'AFD à noyau. Comme la méthode ACP à noyau, la méthode AFD à noyau est une AFD effectuée dans l'espace de caractéristiques k

Le principal inconvénient de ces méthodes est leur temps de calcul. Une méthode d'extraction d'attributs nécessite le calcul des d attributs initiaux pour ensuite extraire les S attributs pertinents (S < d), ces derniers étant obtenus en combinant, linéairement ou non, les d attributs initiaux (voir Figure 3.1). Un autre inconvénient des méthodes d'extraction est qu'elles imposent un effort important à l'utilisateur pour interpréter et comprendre la nouvelle représentation des données : il est difficile de donner une interprétation sémantique des attributs extraits, ces derniers étant une combinaison des attributs initiaux.

3.2.2 Sélection d'attributs

Les méthodes de sélection d'attributs (Yu and Liu, 2003) permettent de sélectionner un sous-ensemble de descripteurs à partir de l'ensemble original selon un critère de performance. Vis-à-vis les méthodes d'extractions d'attributs (voir Figure 3.2), les méthodes de sélection sont caractérisées par leurs temps de calculs minimisés et leurs vitesses de convergence (Koller and Sahami, 1996). En revanche, Les attributs sélectionnés gardent au maximum la sémantique de l'ensemble des données de départ.

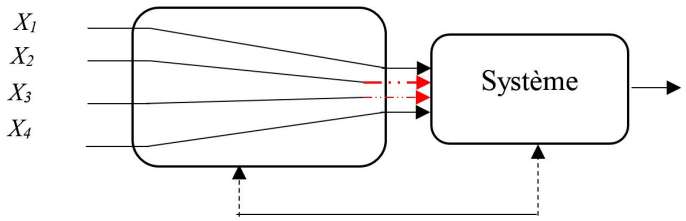


Figure 3.2: Sélection d'Attributs (Guérif, 2006)

Idéalement, les méthodes de sélection d'attributs cherchent le sous ensemble optimal pertinents parmi les 2^N de sous ensemble candidats. Ce qui peut être très couteux en temps de recherche des sous-ensembles exhaustive, donc l'apparition de plusieurs méthodes aléatoires et heuristique qui ont basés sur un critère d'arrêt pour tenter de réduire la complexité. C'est pour cette raison que nous nous sommes intéressée dans ce chapitre aux méthodes de sélection d'attributs et plus particulièrement à aux méthodes wrapper. Ainsi, nous énumérons dans la section 3.2.2.1 les différentes étapes d'un processus de sélection d'attributs. Ensuite, nous introduisons un état de l'art sur les méthodes de sélection d'attributs existantes dans la littérature. Comme ses méthodes donnent une réduction de dimensionnalité mais avec une perte importante d'information, dans la section 2.3 nous introduisons notre algorithme de sélection d'attributs.

3.2.2.1 Étapes de Sélection d'Attributs

La sélection d'attributs pour l'analyse de traces représente une thématique de recherche assez active depuis plusieurs décennies. Elle constitue une étape très importante de prétraitement des traces de grande dimensionnalité.

En effet, l'apparition de grande base de traces dans le domaine d'apprentissage et les systèmes de fouille de données a exigé une réduction de dimension, avant d'entamer la phase de conception du CDSS pour la fouille de traces. La sélection d'attributs cherche à trouver un ensemble de descripteurs pertinents parmi l'ensemble complets et d'éliminer les redondances dans ces derniers. D'où la définition de Féraud et al. (2010),

> "La sélection d'attributs comme étant le processus permettant de choisir un sous ensemble d'attributs pertinents, à partir d'un ensemble d'attributs optimal, selon un certain critère de performance"

En effet, le choix d'un sous ensemble de variable pertinents n'impose pas que chaque variable pertinentes mais l'ensemble de ces variables sont globalement jugés pertinentes et qu'elles sont fortement liés. Par ailleurs, le sous ensemble choisit peut contenir des variables non pertinentes mais qu'elles ont des meilleurs performances avec les autres.

Trois questions importantes à répondre à ce stade avant de définir le processus général d'une sélection d'attributs (Ben Ishak, 2007),

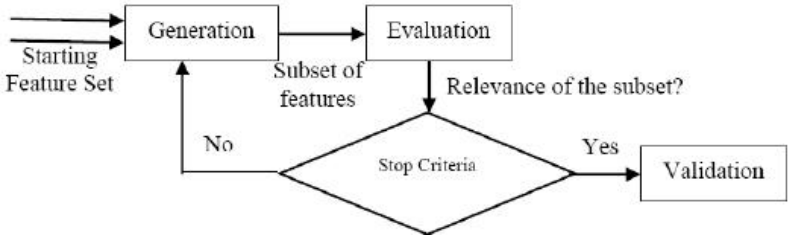


Figure 3.3: Processus de Sélection d'Attributs

- Comment mesurer la pertinence des variables ?
- Comment former le sous-ensemble pertinent ?
- Quel est le critère d'arrêt optimal pour une méthode de sélection d'attributs ?

La pertinence d'une variable peut être définit selon soit son pouvoir discriminant qui produit une meilleure séparation dans les classes dans le cadre d'une méthode supervisée, ou bien son pouvoir prédictif dans le cadre d'une régression. Conditionnellement, il faut déterminer un critère d'arrêt pour évaluer cette pertinence et afin de mesurer la qualité d'une variable ou bien l'ensemble de variable. Par conséquent, un processus de sélection d'attributs peut être représenter par quatre étapes principales illustrés dans la figure 3.3 (Dash et al., 1997):

- Phase de génération
- Phase d'évaluation
- Phase du critère d'arrêt
- Phase de validation

3.2.2.2 Procédure de génération:

La procédure de génération permet, à chaque itération, de générer un sous-ensemble d'attributs qui va être évalué lors de la seconde étape de la procédure de sélection. Cette procédure de génération peut soit commencer avec un ensemble vide d'attributs, soit avec l'ensemble de tous les attributs, soit avec un sous-ensemble d'attributs choisis aléatoirement.

Dans les deux premiers cas, les attributs sont itérativement ajoutés (Forward selection) ou retirés (Backward selection). Dans le troisième cas, soit on ajoute, ou on retire des attributs comme dans les deux premiers cas, soit un nouveau sous-ensemble d'attributs est créé de manière aléatoire à chaque itération (Random generation). Trois grandes approches de génération ont été proposées dans la littérature, la génération complète, la génération aléatoire et la génération séquentielle (Blansché, 2006; Dash and Liu, 1997).

En effet, pour un ensemble d'attributs initial de dimension d, le nombre total de sous- ensembles candidats qui peuvent être générés par la procédure de génération est 2^d. Ce nombre est généralement très élevé surtout lorsque le nombre d d'attributs est élevé.

Génération complète: Dans la procédure de génération complète, une recherche complète du sous-ensemble d'attributs optimal au sens de la fonction d'évaluation utilisée est effectuée. Une recherche exhaustive est complète, cependant la recherche ne doit pas être exhaustive pour qu'elle soit complète.

C'est pour cela, au lieu d'évaluer les 2^d sous-ensemble candidats, différentes fonctions heuristiques peuvent être utilisées afin de réduire l'espace de recherche sans pour autant compromettre les chances de trouver le sous-ensemble optimal. Il s'agit d'utiliser un processus de backtracking permettant de revenir en arrière si la sélection s'engage dans une mauvaise direction de génération.

Génération aléatoire: Les procédures de génération aléatoire parcourent au hasard l'ensemble des 2^d sous- ensembles candidats, le sous-ensemble courant n'est alors pas issu d'une augmentation ou diminution d'attributs du sous-ensemble précédent. Cela permet de ne pas arrêter la recherche lorsque la fonction d'évaluation d'un sous-ensemble atteint un optimum local. Cependant, les 2^d sous-ensembles candidats ne sont pas tous évalués, contrairement aux procédures de génération complète. Un nombre maximal d'itérations est imposé afin que les temps de calcul restent raisonnables.

Les algorithmes génétiques (AG), initiés par Holland en 1975 (Holland, 1992), sont les méthodes de génération aléatoire les plus couramment utilisées (Goldberg et al., 1994). L'avantage de la procédure de génération aléatoire est qu'elle ne nécessite pas l'utilisation de fonction d'évaluation monotone. D'autre part, contrairement aux méthodes de génération complète dont la complexité est exponentielle vis-à-vis de la dimension initiale d de l'espace d'attributs, la complexité de calcul des méthodes basées sur une génération aléatoire est quadratique (Dash et al., 2000; Kudo and Sklansky, 2000). C'est également le cas des méthodes de sélection basées sur les procédures de génération séquentielle.

Génération séquentielle: Le principe des procédures de génération séquentielle est d'ajouter ou de supprimer un ou plusieurs attributs au fur et à mesure des itérations. On distingue alors deux approches de génération séquentielle:

- L'approche de type Forward ou Ascendante : cette approche part d'un ensemble vide d'attributs auquel, à chaque itération sont ajoutés un ou plusieurs attributs.
- L'approche de type Backward ou Descendante : c'est l'approche inverse, elle part de l'ensemble total des attributs. Chaque itération permet de supprimer un ou plusieurs attributs.

Les algorithmes utilisant ces approches de génération sont connues par leur simplicité de mise en œuvre et leur rapidité. Cependant, comme ils n'explorent pas tous les sous-ensembles possibles d'attributs et ne permettent pas de retour arrière pendant la recherche; ils sont donc sous-optimaux. Il est alors possible d'ajouter (ou de retirer) itérativement les attributs. C'est notamment le cas de l'algorithme plus l-take away r. Cet algorithme consiste tout d'abord à élargir le sous-ensemble d'attributs en répétant l fois la procédure Forward, puis à éliminer des attributs en répétant r fois la procédure Backward. Notons que le choix des paramètres l et r influe sur la qualité des résultats ainsi que sur le temps de calcul.

Cet algorithme est très performant lorsque l'on connait a priori la dimension du sous- espace discriminant, mais il reste cependant très couteux en temps de calcul. Pour réduire les coûts de calcul, tout en tentant de conserver un niveau élevé de performance, on peut utiliser les méthodes flottantes, qui sont une extension de l'algorithme plus l-take away r.

L'algorithme SFFS (Sequential Forward Floating Selection) consiste à appliquer après chaque étape Forward des étapes Backward tant que le sous-espace d'attributs correspondant améliore la fonction d'évaluation. L'algorithme SBFS (Sequential Backward Floating Selection) applique le même principe à la différence que les deux étapes sont inversées.

3.2.2.3 Phase d'évaluation

La fonction évaluation permet d'évaluer les attributs ou les sous-ensembles d'attributs générés à l'étape précédente. Elle est utilisée pour mesurer : la pertinence des attributs en les appréciant de manière individuelle, lorsqu'on utilise un algorithme de sélection par classement des attributs et la pertinence des sous-ensembles d'attributs générés par l'une des différentes méthodes de génération présentées ci-dessus, lorsqu'un algorithme de recherche de sous-ensembles est utilisé.

En effet, la sélection d'un sous-ensemble d'attributs optimal est toujours relative au critère utilisé car différents critères ne permettent pas de sélectionner le même sous-ensemble d'attributs optimal. Différentes fonctions d'évaluation ont été proposées pour évaluer un attribut ou un sous-ensemble d'attributs dans un contexte de sélection. Elles peuvent être classées en cinq approches distinctes (Dash and Liu, 1997) :

- Les mesures d'erreur de classification, l'attribut ou les sous-ensembles d'attributs considérés sont évalués en fonction de la qualité de la classification obtenue en utilisant ces attributs. Le sous-ensemble d'attributs le plus discriminant est celui pour lequel le taux d'erreur de classification est le plus faible (Diday, 1982).
- Les mesures d'information, les mesures d'information déterminent le gain d'information pour un attribut considéré, le gain d'information apporté par un attribut étant estimé à partir des probabilités a posteriori. Un attribut f_r est préféré à un attribut f_y si le gain d'information apporté par l'attribut f_r est plus grand que celui apporté par l'attribut f_y (Cover, 1991).

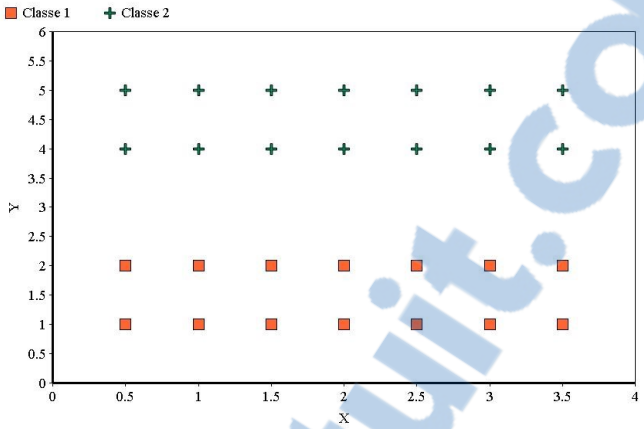


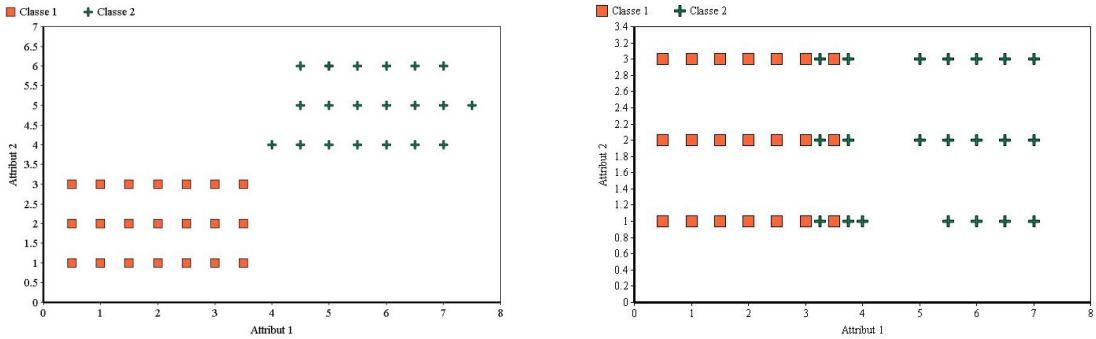
Figure 3.4: Représentation d'un attribut consistant (y) et d'un attribut non consistant (x).

- Les mesures de consistance, les mesures de consistance cherchent à évaluer si l'attribut (ou le sous-ensemble d'attributs) étudié contient les informations nécessaires à la discrimination des classes [Sem04].La figure 3.4 illustre un attribut consistant (f) et un attribut non consistant $(f\backslash)$: on voit aisément dans cette figure que contrairement à l'attribut (f), l'attribut $(f\backslash)$ permet de discriminer les deux classes en présence.

- Les mesures de dépendances, peuvent être divisées en deux catégories : la première est une mesure de corrélation qui quantifie la dépendance des attributs les uns par rapport aux autres. La deuxième catégorie est une mesure de dépendance qui caractérise la corrélation entre un attribut ou un sous ensemble d'attributs et une classe (Dash and Liu, 1997).

- Les mesures de distance, sont aussi appelées mesures de séparabilité, divergence ou de discrimination. Un attribut ou un sous-ensemble d'attributs est sélectionné s'il permet une meilleure séparabilité et cohérence des classes, en effet, le but est de maximiser la dispersion interclasse, afin que les différents points représentatifs soient bien séparés et minimiser la dispersion intra-classe pour que les points de chaque classe soient fortement corrélés.

La figure 3.5 illustre un ensemble de points provenant de deux classes représentées dans deux sous-espaces différents. Dans le premier sous-ensemble formé des attributs f et $f\backslash$, les points représentatifs correspondent aux deux classes sont compacts et séparés (Figure 3.5a), tandis que dans le second sous-ensemble formé des attributs $f\backslash\backslash$ et $f\backslash\backslash\backslash$, ces points représentatifs sont proches et étendus (Figure 3.5b). Le premier sous-ensemble et plus séparable vis-à-vis des deux classes que le second sous-espace.



(a) Espace discriminant d'attributs 1 et 2

(b) Espace non discriminant d'attributs 1 et 2

Figure 3.5: Mesures de discrimination

3.2.3 Méthodes d'extractions de traces pertinentes

Afin de mesurer l'efficacité dans la sélection de la pertinence des traces, un ensemble de méthodes de sélection est présenté. Selon le critère d'évaluation utilisé dans le processus de sélection d'attributs, nous pouvons distinguer entre les approches de type « wrapper », les approches de type « filter » et les approches « Hybrides ». Dans cette section, on va introduire brièvement ces méthodes (Bolón-Canedo et al., 2013).

3.2.3.1 Méthodes Filter

les approches « filter » utilisent une fonction dévaluation basée sur les caractéristiques de l'ensemble des données, indépendamment de tout algorithme de classification, afin de sélectionner certains attributs ou sous-ensembles d'attributs. Ces méthodes sont, rapides plus générales et moins couteuses en temps de calcul, ce qui leur permet d'opérer plus facilement avec des bases de données de très grandes dimensions. Cependant, comme elles sont indépendantes de la méthode de classification, elles ne permettent pas de garantir que le meilleur taux de classification soit obtenu dans l'espace retenu. Selon (Vergara and Estévez, 2014), une approche de type filter doit considérer trois concepts majeurs (Figure 3.6):

- La pertinence est la valeur moyenne de l'information fournit par un seul ou sous-ensemble d'attributs
- La redondance est la dépendance entre deux ou plusieurs attributs, par exemple, Yu and Liu (2004) ont proposés une méthode pour ranger les attributs. Cependant, le sous ensemble d'attribut optimal est composé par les attributs les plus pertinents et les moins pertinents mais qui sont pas redondant.
- Selon le concept de complémentarité qui est définit par le degré d'interaction entre un at-

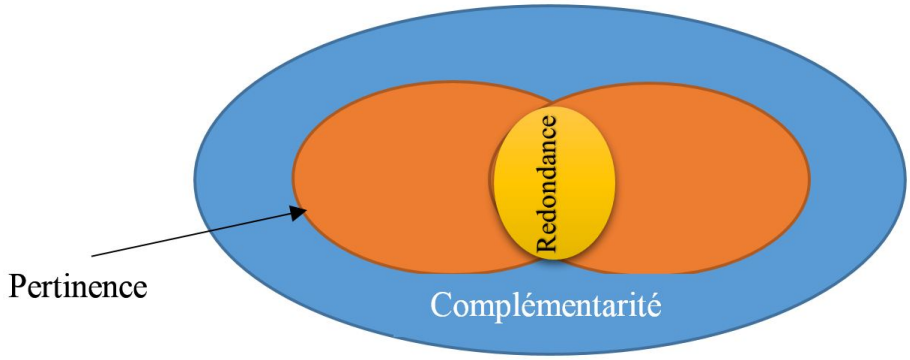


Figure 3.6: Diagramme montrant les différentes relations entre Complémentarité, redondance et pertinence selon (Vergara and Estévez, 2014)

tribut et l'ensemble des autres attributs. Un prototype unifié basé sur l'information mutuel a été proposé afin d'indiquer les limites différentes méthodes de sélection d'attributs basés sur l'information mutuelle.

Correlation-based feature selection (CBFS) , est un algorithme de filtrage simple basé sur une fonction d'évaluation de corrélation (Hall, 1999). Afin de sélectionner le meilleur sous-ensemble de fonctionnalités, le CBFS sélectionne les attributs les plus corrélés avec l'attribut de classe; les attributs redondants sont éliminés des fonctions s'ils sont corrélées l'un avec l'autre.

Consistency-based filter (CBF) , les traces sont projetées sur le sous-ensemble sélectionné des attributs et l'évaluation est basée sur le niveau de cohérence au sein des valeurs de la classe (Dash and Liu, 2003).

INTERACT , est un algorithme de filtre (Zhao and Liu, 2007). L'algorithme se compose de deux grandes étapes; les attributs sont d'abord individuellement classés par ordre décroissant utilisant l'incertitude symétrique (SU). Dans la deuxième étape, les descripteurs sont évaluées un par un selon l'ordre déjà préétablie. Le descripteur est conservé si sa contribution dans la cohérence est plus grande qu'un seuil établi avec la classe.

ReliefF (Robnik-Šikonja and Kononenko, 2003), est un algorithme robuste qui peut traiter les problèmes multi-classe. Similaire à l'algorithme de Relief, ReliefF estime la qualité des attributs en fonction de la façon dont leurs valeurs peuvent séparer les cas similaires. Un descripteur de trace est pertinent s'il est capable de différencier les instances de différentes classes et avoir la même valeur pour les instances de la même classe (Figure 3.7).

— Set all weights ω ←(0,0,…….0).
— **for** t=1 to N **do**
— select a random instance x from S.
— find k nearest hits H_t.
— **for** each class C≠ class(x) **do**
— from class C find k nearest misses M_t.
— **end for**
— **for** feature i=1 to M **do**
— $\omega_i \leftarrow \omega_i + \sum_{C \neq class(x)} \frac{P(c)}{1-P(Class(x))} \frac{\sum_{\forall \bar{x} \in M_t} \|x_i - \bar{x}_i\|^2}{k} - \frac{\sum_{\forall \bar{x} \in \bar{H}_t} \|x_i - \bar{x}_i\|^2}{k}$
— **end for**
— **end for**

Figure 3.7: L'algorithme ReliefF

3.2.3.2 Méthodes Wrapper

les approches "wrappers" utilisent le taux d'erreur de classification comme critère d'évaluation (mesures d'erreur de classification) (Kohavi and John, 1997). Ils incorporent alors l'algorithme de classification dans la procédure de recherche et sélection d'attributs. Ces méthodes permettent d'obtenir de bonnes performances. Cependant, l'utilisation de telles méthodes nécessite pour chaque sous espace d'attributs candidats d'effectuer la classification, ce qui peut devenir coûteux en temps de calcul surtout lorsque la dimension d de l'espace d'entrée est grande. De même, ces méthodes sont très dépendantes de l'algorithme de classification utilisé comme critère d'évaluation. Ce dernier, montre un avantage de ces méthodes car ces méthodes sont généralement des méthodes de classification donc elles peuvent être adaptés dans n'importe qu'elle problème de classification et d'analyses de traces. A cause de leur efficacité et leur dépendance de tout algorithme de classification, les approches de type "wrapper" sont plus répandues et d'emploi courant. Cependant, ces approches sont combinées avec d'autres méthodes de classification. La première technique de FS la plus reconnu est Best First Search (BFS) comme une méthode de génération. BFS est un algorithme de recherche qui explore un graphe en élargissant les nœuds les plus pertinentes ayant le meilleur score donné par la mesure d'évaluation (Dechter and Pearl, 1985).

BFS commence d'abord avec un ensemble vide de descripteurs et génère tous les les extensions possibles en rajoutant un descripteur. Le sous-ensemble avec la plus haute évaluation est sélectionné et élargi de manière similaire en ajoutant un descripteur simple. Si le sous-ensemble sélectionné des attributs n'en résulte aucune amélioration, la recherche revient au sous-ensemble de descripteurs précédant. Un BFS explorera tout l'espace des attributs. Le meilleur sous-ensemble trouvé est retourné lorsque la recherche se termine. La figure 3.8 montre l'algorithme BFS.

1. Put the initial state on the OPEN list, CLOSED list → ∅, BEST list → *initial*.
2. Let $\alpha = argmax_{x \in OPEN} f(x)$
3. Remove α from OPEN, add α to CLOSED
4. If $f(\alpha) - \varepsilon > f(\text{BEST})$, *then* $BEST \leftarrow \alpha$
5. Expand α: *apply all operators to* α, *giving* α*'s children*
6. For each child not in the CLOSED or OPEN list, evaluate and add to the OPEN list,
7. If BEST changed in the last k expansions, go to 2
8. Return BEST.

Figure 3.8: L'algorithme BFS

la deuxième méthode wrapper la plus répandue produit un sous-ensemble de descripteurs de traces pertinents en utilisant la méthode sequential forward search (SFS) (Schaffernicht et al., 2007). Le but de cette méthode est d'ajouter un ou plusieurs attributs de plus en plus. Comme l'algorithme BFS, l'algorithme SFS commence d'abord avec un ensemble vide. Pour chaque itération, un descripteur est ajouté au sous-ensemble des descripteurs sélectionnés et ensuite évalués. Si le descripteur supplémentaire montre une amélioration de la fonction de l'évaluation du sous-ensemble résultant. Le processus se poursuit jusqu'à absence d'amélioration ou un nombre prédéfini nombre utilisateur fixe de caractéristiques est atteint.

- WrapperSubsetEval (Witten and Frank, 2005) a été utilisé pour évaluer les sous-ensembles de descripteurs sélectionné en utilisant un algorithme de classification (voir section 2.3.2).

3.2.3.3 Méthodes Hybrides

Pour combiner les avantages des deux méthodes, des algorithmes hybrides "Hybride" ont été proposés. Le processus de sélection d'attributs est effectué conjointement au processus de classification. Une fonction d'évaluation de type "filter" est tout d'abord utilisée pour présélectionner les sous-espaces d'attributs les plus discriminants. Puis les taux d'erreurs de classification obtenus en considérant chaque sous-espace discriminant précédemment sélectionné sont comparés afin de déterminer le sous espace final (Jin et al., 2012).

Beaucoup de travaux ont été mis, dans ce sens. Par exemple, dans les travaux de (Korfiatis et al., 2013), une approche Hybride à deux phases a été proposée pour l'analyse des traces d'un ensemble de patients de polycythaemia. Dans la première phase le meilleur sous ensemble d'attributs de taille S fixé apriori était sélectionné utilisant une approche wrapper. Ensuite, les auteurs ont variés la taille de l'ensemble d'attributs sélectionné S afin d'utiliser une approche filter pour la sélection de nouveaux attributs qui vont améliorer la performance du classificateur.

L'approche dans (Bermejo et al., 2012) combine entre une approche filter base sur l'information mutuelle et une approche wrapper qui utilise les attributs rangés par la première approche pour la classification ce qui va impliquer une réduction en temps de calculs pour la méthode wrapper.

(Oreski and Oreski, 2014) ont proposé un systeme pour l'évaluation des risques de crédit bancaire. Cette méthode utilise le même principe que la méthode proposé dans (Bermejo et al., 2012). Pour la méthode wrapper, les auteurs ont utilisé l'algorithme génétique pour l'exploration de l'espace d'attributs et le réseau neurone pour la classification.

Méthodes Hybrides effectuent la sélection en phase d'apprentissage du processus de classification de l'algorithme de fouille de données. Afin de montrer l'efficacité de notre contribution, nous avons également comparé nos résultats avec différentes méthodes Hybrides [première approche hybride et la deuxième approche hybride (Ruiz et al., 2012)]: - Feature selection-perceptron (FS-P) l'idée de base de cette méthode est de passer par une phase d'apprentissage utilisant l'algorithme du perceptron multicouche (MLP) (voir section 4) et puis les poids d'interconnexion sont utilisés pour produire classement des attributs (Mejía-Lavalle et al., 2006).

3.2.3.4 Critère d'arrêt

Le nombre optimal d'attributs n'étant pas connu a priori, il sera fixé grâce à un critère d'arrêt du processus de sélection. L'utilisation d'une règle pour contrôler la procédure de sélection permet d'arrêter la recherche lorsqu'aucun nouvel attribut n'est suffisamment informatif. C'est un choix souvent défini en fonction de la procédure de recherche (Sémani-Delmi, 2004) et/ou du critère d'évaluation (Pietra et al., 1997).

Les critères d'arrêt les plus fréquents sont:

- basés sur l'algorithme de génération (Zhu et al., 2007) : on peut par exemple décider d'arrêter la recherche en fixant un seuil sur le nombre d'attributs à sélectionner ou sur le nombre d'itérations. Cependant, dans de nombreuses applications, le nombre d'attributs à sélectionner est très difficile à fixer au préalable. De même, un critère fondé sur un nombre maximal d'itérations peut s'avérer brutal et arrêter trop tôt ou trop tard la sélection (Sémani-Delmi, 2004).

- basés sur l'évaluation (Pietra et al., 1997) : dans ce cas, on arrête la recherche en fixant un seuil soit sur la fonction d'évaluation, soit sur la différence entre la valeur d'évaluation à l'étape d et la valeur d'évaluation à l'étape $d-1$, c'est-à-dire lorsque l'ajout ou la suppression d'un attribut n'apporte pas un gain de discrimination suffisant. Par exemple, lorsque l'approche "wrapper" ou l'approche "Hybride" est utilisée, les taux de bonne classification obtenus par les différents sous-espaces sont comparés pour mesurer le gain d'information. On peut ainsi décider d'arrêter la procédure de sélection dès que ce taux diminue ou alors dès qu'il atteint un certain seuil.

Le choix du critère d'arrêt est ainsi un choix délicat qui reste un problème non résolu dans de nombreux algorithmes de sélection. On ne connait pas à l'avance la dimension bd du sous-espace sélectionné.

3.2.3.5 Phase de validation

La validation ne fait pas partie de la procédure de sélection d'attributs mais elle permet de tester la validité du sous-ensemble d'attributs sélectionnés en effectuant plusieurs tests sur des exemples de données générées artificiellement et/ou sur des données réelles.

L'ensemble des données est généralement divisé en deux sous-ensembles distincts : le sous ensemble d'apprentissage constitué des prototypes des classes (données avec leurs labels) et le sous-ensemble de test dont on ne connait pas les labels de classes de ses données. Selon la répartition des données entre ces deux sous-ensembles, il existe différentes approches de validation :

- La méthode Holdout: les données sont réparties en deux parties : le sous ensemble d'apprentissage et le sous ensemble de test.
- La méthode de resubstitution, le même ensemble d'apprentissage est utilisé pour le test, cette méthode est très rarement utilisée.
- La validation K-Croisée : dans cette méthode l'ensemble de données et fragmenter en K ensembles de taille à peu près égales, pour chaque itération on prend un sous ensemble pour la validation et les $k-1$ sous ensemble pour l'apprentissage. Le processus se terminera lorsque tous les sous-ensembles ont étaient validés. C'est la méthode la plus utilisées.

3.3 Conclusion

La sélection d'attributs est un axe de recherché très actifs et productif pour la réduction de dimension des traces. Il était démontré l'efficacité de ses techniques pour améliorer les performances des méthodes de classification. Cependant, la sélection d'attribut est une tache non triviale, dans ce chapitre on a introduit la sélection d'attribut pour la pertinence des traces. Le processus général a été présenté dirigée par l'étude des différentes méthodes existant dans la littérature.

Enfin, dans le chapitre 4, on va présenter un ensemble d'expérience. Nous allons montrer à travers que notre méthode produit les meilleurs résultats vis-à-vis les différentes méthodes de la littérature utilisant plusieurs data sets (Mokeddem S et al., 2016).

Chapitre 4

Une Méthode Effective pour la Sélection d'Attribut basée Algorithme Génétique Wrapper Naïve Bayes

Sommaire

4.1 Introduction

Nous avons vu dans le chapitre 3 que la phase d'extraction de la pertinence pour l'analyse des traces est une étape importante pour l'aide à la décision. Le but des travaux d'analyse de la pertinence est de développer des méthodes et des outils visant à extraire les informations pertinentes à partir des traces. Nous avons choisis une variété de domaines pour évaluer notre méthode proposée. En outre, nous avons aussi focalisés sur les traces des patients du CAD.

Une quantité importante d'information concernant les différentes traces pour différents domaine ont été incluses dans notre étude. Ces traces représentent une source de connaissances très utiles dans plusieurs applications de prise de décision, la prévention et la fouille de données. Notre intérêt qui porte sur l'analyse de la pertinence des traces a été principalement motivé par la conviction qu'elle peut être appliquée dans les différents domaines et plus précisément la littérature médicale, en particulier les traces des patients qui intègrent un grand nombre d'informations.

A la lumière de l'approche proposée au chapitre précédent, il est de situer notre approche par rapport aux familles de méthodes présentées dans la littérature et donc notre apport. Dans ce chapitre on va évaluer notre approche vis-à-vis les différentes méthodes. Cependant, nous décrirons les données et les mesures utilisées pour évaluer notre approche et examiner les résultats des expériences. Dans la section qui suit on va décrire les différentes bases de traces utilisées. La section 3 présente les méthodes et les mesures utilisées lors des expérimentations. Enfin, dans la section 4 nous montrerons les différents scénarios expérimentaux avec une discussion des différents résultats.

4.2 Une méthode effective pour la sélection d'attribut basée Algorithme Génétique wrapper Naïve Bayes

Dans cette partie, on va présenter notre contribution dans les méthodes de sélection de variable. C'est pour cela, on va focaliser sur les différents matériaux utilisés lors de la conception de notre algorithme. Notre algorithme hybride entre l'algorithme génétique et la méthode Naïve Bayes pour sélectionner les variables les plus pertinents (?). En effet, la motivation d'utiliser l'algorithme génétique est son aptitude de ne peux pas explorer toute les régions ce qui est intéressent en matière de couts de calculs et de temps. Notre algorithme considère deux étapes : dans la première étape un ensemble d'attribut est sélectionnée, la dimension est réduite utilisant l'algorithme génétique avec la meilleur configuration de ce dernier, après on fait appel à la méthode Naïve Bayes de classification pour évaluer et mesurer la précision de l'algorithme. L'approche va être détaillée dans les sections qui suivent.

4.2.1 Algorithme Génétique Pour la sélection de variable et algorithme proposé

L'algorithme génétique (AG) est un algorithme de recherche basé sur les mécanismes de la sélection naturelle et de la génétique proposé par (Goldberg et al., 1994). Il combine une stratégie de survie des plus forts avec un échange d'information aléatoire mais structuré. Pour un problème pour lequel une solution est inconnue, un ensemble de solutions possibles est créée aléatoirement, cet ensemble est appelé population.

Les AGs ont été initialement développés par Holland (1992). C'est au livre de Golberg (1989) que nous devons leur popularisation. Leurs champs d'application sont très vastes. Outre l'économie, ils sont utilisés pour l'optimisation de fonctions Jong (1980), en finance (Pereira, 2000), en théorie du contrôle optimal (Krishnakumar and Goldberg, 1992; Marco et al., 1996; Michalewicz et al., 1992),ou encore en théorie des jeux répétés (Axelrod, 1987) et différentiels (Özyildirim (1996, 1997) et (Alemdar and Özyildirim, 1998; Özyildirim, 1996, 1997). La raison de ce grand nombre d'application est claire : simplicité et efficacité. Bien sûr d'autres techniques d'exploration stochastique existent, la plus connue étant le recuit simulé (simulated nnealing) pour une association des deux méthodes (Davis, 1987).

Pour résumer, Lerman and Ngouenet (1995) distinguent 4 principaux points qui font la différence fondamentale entre ces algorithmes et les autres méthodes :

1. Les algorithmes génétiques utilisent un codage des paramètres, et non les paramètres eux-mêmes.

2. Les algorithmes génétiques travaillent sur une population de points, au lieu d'un point unique.

3. Les algorithmes génétiques n'utilisent que les valeurs de la fonction étudiée, pas sa dérivée, ou une autre connaissance auxiliaire.

4. Les algorithmes génétiques utilisent des règles de transition probabilistes, et non déterministes.

Les caractéristiques (dans le cas de la sélection d'attributs « les variables ») sont alors utilisées dans des séquences de gènes qui seront combiner entre eux afin de produire de nouvelles séquences de gènes ou chromosomes. Chaque solution est liée à un chromosome ou bien individu, cet individu est évalué et classer selon un critère de qualité avec la meilleur solution trouvée. Comme dans l'évolution biologique, chacune des meilleures solutions de la population va transmettre son héritage et produisent des solutions aussi bien meilleures que celles d'eux. Une nouvelle génération et donc créer en appliquant certains opérateurs génétiques sur les parents. Par conséquent, la génération produite possède les meilleurs caractéristiques de leurs deux parents donc une meilleure solution au problème. Ce processus est répété plusieurs fois jusqu'à ce que tous les individus possèdent le même héritage génétique. Les individus de la dernière génération qui

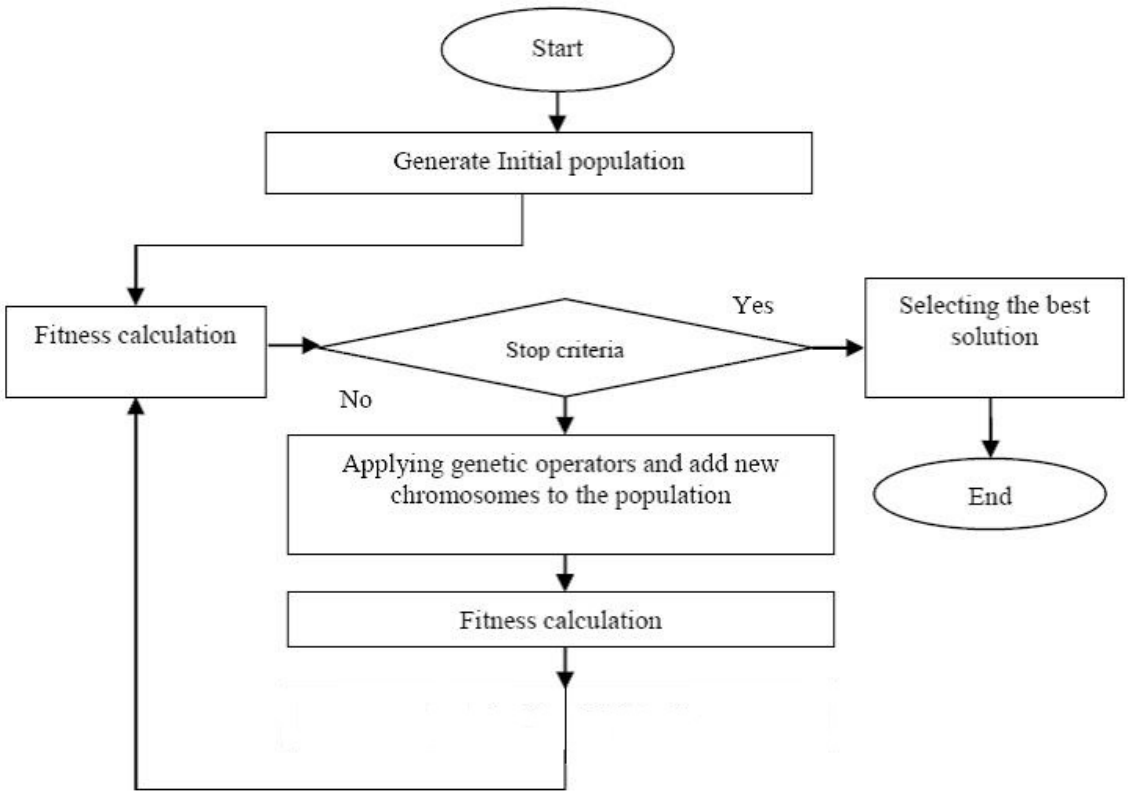


Figure 4.1: Description générale d'un algorithme génétique

possèdent des gènes différents de ceux de leurs ancêtres sont généralement de meilleures qualités donc c'est la meilleure solution au problème. L'algorithme génétique de base comporte trois opérations simples de base:

- Sélection
- Croisement
- Mutation

Figure 4.1 illustre le diagramme générique d'un algorithme génétique pour la sélection d'attributs. Selon Lerman and Ngouenet (1995) un algorithme génétique est défini par :

1. Individu/chromosome/séquence : une solution potentielle du problème
2. Population : un ensemble de chromosomes ou de points de l'espace de recherche
3. Environnement : l'espace de recherche
4. Fonction de fitness : la fonction (positive) que nous cherchons à maximiser

Avant d'expliquer le processus général d'un algorithme génétique, il faut tout d'abord expliquer définir quelque termes importants généralement définit sous l'hypothèse encodage binaire.

Définition 4.2.1 (séquence/chromosome/individu) Nous appelons un chromosome (séquence/individu) A de longueur $l(A) = n$ toutes séquences $A = (a_1, a_2, a_3, ..., a_n)$, tel que $\forall i \in 1, ..., n, a_i \in B0, 1$

L'encodage des individus. L'encodage normal d'un individu ayant plusieurs paramètres est de coder chaque gène en binaire. Chaque gène est tronqué l'un après l'autre pour former un chromosome. La valeur de chaque gène est appelé allèle. Exemple: soit un chromosome à quatre paramètres a_1, a_2, a_3, a_4 codé à trois bits :

$$a_1 = 010$$

$$a_2 = 001$$

$$a_3 = 100$$

$$a_4 = 110$$

Le chromosome approprié à ces paramètres serait le suivant:

$$y = \overbrace{010}^{a_1}\overbrace{001}^{a_2}\overbrace{100}^{a_3}\overbrace{110}^{a_4}$$

Un chiffre code en binaire dans ce cas ne représente pas une valeur spécifique mais plutôt un intervalle. Par exemple, si on code des valeurs de 0 à 1 à l'aide de 5 bits, le chiffre 11111 représente l'intervalle de $31/32$ à $32/32$, plutôt que la valeur de $31/32$ exactement.

Définition 4.2.2 (Fitness d'un chromosome) Nous appelons fitness d'un individu a toute valeur positive $f(a) > 0$, ou f est précisément appelée fonction de fitness

L'objectif d'un GA est alors simplement de trouver le chromosome qui maximise cette fonction f. Bien évidemment, chaque problème particulier nécessitera ses propres fonctions d et f . Dans notre cas, c'est celui qui maximise la performance de classification.

Les AGs sont alors fondés sur les phases suivantes (voir figure 4.1) :

- Initialisation. Une population initiale de N individus est tirée aléatoirement.
- Évaluation. Chaque individu est décodé, puis évalué.
- Sélection. Création d'une nouvelle population de N individus par l'utilisation d'une méthode de sélection.

- Reproduction. Possibilité de croisement et mutation au sein de la nouvelle population.
- Retour à la phase d'évaluation jusqu'à l'arrêt de l'algorithme.

4.2.1.1 Initialisation

Pour ce qui est de la phase d'initialisation, la méthode est assez élémentaire. On effectue un tirage aléatoire de N chromosomes (population initiale) dans l'espace des chromosomes permis. Selon le codage binaire, et la taille l de la chaîne, nous effectuons pour un chromosome l tirage dans 0,1 avec équiprobabilité.

4.2.1.2 Opérateurs

Les opérateurs ont un rôle principal dans la performance d'un AG. Nous en citons les trois principaux : l'opérateur de sélection, de croisement et de mutation. Si le fondement de chacun de ces opérateurs est facilement compréhensible, il est cependant difficile d'étaler l'importance isolée de chacun de ces opérateurs dans la performance de l'AG. Cela tient pour partie au fait que chacun de ces opérateurs agit selon divers critères qui lui sont propres (valeur sélective des individus, probabilité d'activation de l'opérateur, etc.).

Opérateur de Sélection, Cet opérateur est peut-être le plus important puisqu'il permet aux individus d'une population de survivre, de se reproduire ou de mourir. En règle générale, la probabilité de survie d'un individu sera directement reliée à son efficacité relative au sein de la population.

Il existe nombreuses méthodes pour la reproduction. La méthode la plus connue et utilisée est, la roue de loterie biaisée (roulette wheel) de Golberg (1989). Selon cette méthode, chaque séquence sera dupliquée dans une nouvelle population proportionnellement à sa valeur d'adaptation. On effectue, en quelque sorte, autant de tirages avec remises qu'il y a d'éléments dans la population. Or, dans le cas d'un codage binaire, la fitness d'une séquence particulière étant $f(p(c_i))$, la probabilité avec dont il sera réintroduit dans la nouvelle population de taille n est:

$$\frac{f(p(c_i))}{\sum_{i=1}^{N} f(p(c_i))}$$

Les chromosomes ayant une meilleure fitness ont donc plus de chance d'être choisis. On parle alors de sélection proportionnelle. Le désagrément capital de cette méthode repose sur le fait qu'un chromosome n'étant pas le meilleur peut tout de même imposer la sélection. Elle peut aussi produire une perte de différence par la domination d'un super chromosome. Une autre difficulté est sa faible performance vers la fin quand l'ensemble des chromosomes se ressemblent. (Dawid, 2011) récapitule très bien tous ces inconvénients :

Pensez à une situation où une chaîne [chromosome pour nous] de la population a comparativement une fitness élevée mais n'est pas optimal ou proche de l'optimum. Disons que la fitness de cette chaîne est dix fois plus grande que la fitness moyenne. Il pourrait facilement arriver, après quelques générations, que la population ne soit entièrement constituée que de cette chaîne. Dans une telle situation, GA ne s'améliore plus en matière de fitness et l'optimum ne sera pas trouvé. Ce phénomène est nommé "convergence prématurée" et est l'un des complications les plus fréquents lors de l'utilisation des GAs. Un autre problème originaire de la sélection proportionnelle est celui du "fine tuning" à la fin de la recherche.

Une solution alternative à ce problème est l'utilisation d'une fonction fitness modifiée et non pas l'utilisation d'une autre méthode de sélection. De ce fait, nous pouvons user une transmutation d'échelle (scaling) afin de réduire ou étendre de manière artificielle la distance relative entre les fitness des chromosomes.

Cependant, il existe d'autres méthodes, la plus répandue étant celle du tournoi (tournament selection): dans cette méthode on choisit deux chromosomes aléatoirement dans la population et on introduit le meilleur des deux dans la nouvelle population. Ce processus est refait jusqu'à ce que la nouvelle population soit complète. Cette méthode produit de meilleurs résultats. Néanmoins, aussi important que soit la phase de sélection, elle ne génère pas de nouveaux chromosomes dans la population. Ceci est le rôle des autres opérateurs de croisement et de mutation.

Opérateur de Croisement L'opérateur de croisement permet la création de nouveaux individus selon un processus fort simple. Permet donc l'échange d'information entre les chromosomes (individus). Tout d'abord, deux individus, qui forment alors un couple, sont tirés au sein de la nouvelle population issue de la reproduction. Puis un (potentiellement plusieurs) site de croisement est tiré aléatoirement (chiffre entre 1 et $l-1$). Enfin, selon une probabilité p_c que le croisement effectue, les segments finaux (dans le cas d'un seul site de croisement) des deux parents sont alors échangés autour de ce site.

Cet opérateur permet la création de deux nouveaux individus. Toutefois, un individu sélectionné lors de la reproduction ne subit pas nécessairement l'action d'un croisement. Ce dernier ne s'effectue qu'avec une certaine probabilité. Plus cette probabilité est élevée et plus la population subira de changement. Quoi qu'il en soit, il se peut que l'action conjointe de la reproduction et du croisement soit insuffisante pour assurer la réussite de l'AG. Ainsi, dans le cas du codage binaire, certaines informations (i.e. caractères de l'alphabet) peuvent disparaître de la population. Ainsi aucun individu de la population initiale ne contient de 1 en dernière position de la chaîne, et que ce 1 fasse partie de la chaîne optimale à trouver, tous les croisements possibles ne permettront pas de faire apparaître ce 1 initialement inconnue. En codage réel, une telle situation peut arriver si utilisant un opérateur simple de croisement, il se trouvait qu'initialement toute la population soit comprise entre 0 et 40 et que la valeur optimale était de 50. Toutes les combinaisons convexes possibles de chiffres appartenant à l'intervalle $[0; 40]$ ne permettront jamais d'aboutir

à un chiffre de 50. C'est pour remédier entre autre à ce problème que l'opérateur de mutation est utilisé. Exemple : Soit $k = 2$ pour deux parents c_1 et c_2 codés sur 4 bits donc $l = 4$ les deux fils résultant s_1 et s_2 sont les suivant:

$$c_1 = 0101$$

$$\underbrace{c_2 = 0010}$$

$$s_1 = 0111$$

$$s_2 = 0000$$

Ce sont ces deux opérations, la sélection et la reproduction, qui sont à la base des algorithmes génétiques. Ceci peut paraitre simple à première vue, puisque aucune opération mathématique complexe n'a été effectuée. Mais on peut comparer le processus précédent à l'innovation humaine : souvent, les découvertes n'arrivent pas par chance. Elles sont le résultat d'un échange d'idées qui crée d'autres idées et finalement m'mènent à une solution désirée.

Opérateur de Mutation Le rôle de cet opérateur est de modifier aléatoirement, avec une certaine probabilité, la valeur d'un composant de l'individu. Dans le cas du codage binaire, chaque bit $a_i \in 0, 1$ est remplacé selon une probabilité p_m par son inverse $a_i = 1 - a_i$. Tout comme plusieurs lieux de croisement peuvent être possibles, nous pouvons très bien admettre qu'une même chaîne puisse subir plusieurs mutations.

La mutation est traditionnellement considérée comme un opérateur marginal bien qu'elle confère en quelque sorte aux algorithmes génétiques la propriété d'ergodicité (i.e. tous les points de l'espace de recherche peuvent être atteints). Cet opérateur est donc d'une grande importance. Il a de fait un double rôle: celui d'effectuer une recherche locale et/ou de sortir d'une trappe (recherche éloignée). Exemple: mutation du deuxième bit.

$$c_1 = 1101 \Longrightarrow c_1^{\backprime} = 1001$$

Une explication plus complète de ces phénomènes ainsi qu'une preuve théorique de leur performance sont disponibles dans le livre de Goldberg (Golberg, 1989)

4.2.2 Réduction de dimension pour la classification des traces

Dans notre proposition, on fait appel aux algorithmes génétiques pour le processus de génération (Figure 4.1). GA est une méthode heuristique adapté aux problèmes de recherches basé sur la théorie d'évolution. Pour l'adapter à notre problème de réduction de dimension des traces, l'idée est d'encoder chaque ensemble de variable descriptive en chromosomec_i. Ce groupe de chromosomes, qui consiste à une population, forme l'espace de recherche. Une fonction de fitness a comme rôle d'évaluer la performance des chromosomes (c'est à dire l'ensemble des variables

Paramétrer	Valeurs
Population initiale	Aléatoire
Taille de la population	50
Nombre maximale d'attributs	n/a
Nombre minimale d'attributs	n/a
La mesure d'évaluation	Exactitude de classification
La fonction d'évaluation	Naïve Bayes
Méthode de sélection	loterie biaisée
Sauvegarde des meilleurs chromosomes	Oui
Probabilité de croisement: p_c	0.9
Probabilité de mutation: $p_(m)$	0.05
Critère d'arrêt	Nombre de génération
Nombre de génération	40
Arrêt préalable	Non

Tableau 4.1: Paramètre de l'algorithme proposé

sélectionnées) pour mesurer sa proximité de l'optimum. Par conséquent, dans notre algorithme, on a utilisé Naïve Bayes pour évaluer la performance des chromosomes. Cette fonction fitness, va permettre non seulement d'évaluer la performance des chromosomes, mais elle va donner à notre algorithme l'aptitude d'un classificateur.

La population initiale (chromosomes) est utilisée pour générer de nouveaux sous-ensembles d'attributs (individus) en usant les différents opérateurs génétiques (Sélection, croisement et mutation) pour assurer un passage d'une génération a une autre jusqu'à atteindre le critère d'arrêt. Pour chaque itération, le nombre de chromosomes reste toujours constant avec l'élimination des individus ayant une faible valeur de fitness. Ce processus est réitéré jusqu'à avoir la meilleur fitness ou atteindre un nombre d'itérations maximal fixé au préalable (Yan et al., 2008).

Un sous ensemble d'attributs est encodé en chaine de bits ou chaque bit représente un descripteur. La valeur 1 signifie la sélection du descripteur et la valeur 0 non. Le nombre de gènes contenant dans un chromosome est égale au nombre d'attribut décrivant une trace. La taille de la population consiste du nombre de chromosomes dans l'espace de recherche et on l'a fixé à 50. La population initiale est générée aléatoirement.

On a utilisé la fonction fitness NB_{eval} , généralement cette fonction évalue la pertinence d'un chromosome. Cependant, dans notre algorithme, la pertinence d'un chromosome est représentée en fonction de sa précision de classification. Donc, l'aptitude d'un chromosome d'être élu comme solution optimal est représenté par la précision de classification produite par NB_{eval} pour chaque itération. Pour le processus de génération, on utilise les trois opérateurs génétiques. L'opérateur de sélection dépend de la fonction fitness, car les meilleurs chromosomes sont retenus dans chaque itération en se basant sur la méthode la roue de loterie biaisée. L'opérateur de croisement prend deux points d'échange de valeur entre deux chromosomes différents avec une probabilité de 0.9. Pour l'opérateur de mutation dans notre algorithme est sélectionnée avec

Algorithm 1 Algorithme Proposé

```
Initialize: répartir l'ensemble de données 10 sous echantillons
for k = 1 to 10 do
    test_data_all = 1 sous echantillon pour test NB_wrap
    train_data = 9 sous echantiollon pour l'apprentissage de NB_eval
    for num_generations = 1 to 40 do
        encoder les attributs comme chromosomes binaire
        generer aléatoirement une population de 50 chromosomes
        Evaluer la population initiale en utilisant NB_eval
        Appliquer l'opération de croisement avec une probabilité de 0.9
        Appliquer l'opération de mutation avec une probabilité de 0.05
        Evaluer la nouvelle population générée avec NB_eval et comparer avec 8
        Remplacer les mauvais chromosomes de 7 avec les meilleurs chromosomes de 10
    end for
    Construire un modele NB_wrap utilisant train_data
    Tester NB_wrap utilisant test_data_all
    Calculer la performance pour k
end for
Calculer la performance moyenne pour les 10 sous echantillons
end
```

une probabilité de 0.05. Table 4.1 montre les différents paramètres de notre algorithme.

L'algorithme proposé utilise les différents opérateurs génétiques avec le critère d'arrêt dans le processus de sélection. Tandis, l'algorithme s'arrête lorsque le nombre d'itération num_generation est égal à 40 (L'algorithme 1). L'algorithme est composé de deux boucles imbriquées, la première comprend le processus de génération. Les attributs sont encodé en binaire après la population initiale est aléatoirement généré et évalué utilisant la fonction fitness NB_{eval} . En appliquant les différentes opérateurs génétiques, la nouvelle génération produite est évalué avec NB_{eval}.et par la suite comparé avec la dernière génération produite pour sélectionné les meilleurs chromosomes. Dans la deuxième boucle, l'évaluation de l'ensemble de traces collecté avec le sous ensemble de variables est validé utilisant la validation croisée utilisant NB_{wrap}.

4.3 Data sets

Le corpus d'apprentissage (en anglais, « dataset ») est un élément essentiel à la construction d'un système d'analyse de traces. Plusieurs sites web proposent gratuitement des corpus d'apprentissage et de test bien structurés pour réaliser des travaux portant sur l'analyse des traces de différents domaines.

La performance de l'approche proposée pour l'analyse de la pertinence des traces des patients a été évalué utilisant multiple base de traces issues des différentes base citons parmi: University of California Irvine (UCI)(Asuncion and Newman, 2007) et KEEL (Derrac et al., 2015). L'Annexe

Dépot de données	Data set	Abbreviation	No. d'attributs	No. d'instances	No. de classes
UCI	Arrythmia	ARRH	280	452	16
	Hypothyroid	HYPO	30	3772	4
	Letter	LETT	17	20000	26
	Lymphography	LYMPH	19	148	4
	Multifeat	MULF	217	2000	10
	Mushroom	MUSH	23	8124	2
	Sick	SICK	30	3772	2
	Soybean	SOBE	36	683	2
	Splice	SPLI	62	3190	19
	Waveform	WAVE	41	5000	3
KEEL	Optical Digits	OPTD	64	5620	10
	Colon Cancer	COLC	2000	62	2
	Anneal.Orig	ANNO	39	898	6
	Colic	COLI	23	368	2
	Cylinder-bands	CYLB	40	540	2
	kdd_synthetic_control	KSYC	62	600	6
	Autos	AUTO	26	159	6
	Chess	CHES	37	3196	2
	Coil2000	COIL	86	9822	2
	Connect-4	CONN	43	67557	3
	Dermatology	DERM	35	366	6
	Eastwest	EASW	26	213	2
	Kddcbuf_overflow_v_back	KOVB	42	2233	2
	Kddcup-guess_passwd_v_satan	KGPS	42	1642	2
	Movment_libras	MOVL	91	360	15
	Musk1	MUSK	168	476	2
	Penbased	PENB	17	1100	10
	Sonar	SONA	61	208	2
	Spambase	SPAM	58	4597	2
	Spectfheart	SPEC	45	267	2
	Thyroid	THYR	22	720	3
	Tiger	TIGE	232	12220	2
	Vehicle3	VEHI	19	846	2
	Vowel0	VOWE	14	988	2

Tableau 4.2: Description des différentes bases de traces

A et L'Annexe B décrivent les différentes bases de traces. Table 4.2 montre les différentes caractéristiques des bases de traces.

4.4 Évaluation des algorithmes

L'évaluation est une phase essentielle à tout processus de sélection et de fouille de traces pertinentes. Elle consiste à vérifier que la performance et la qualité de sélection de traces dans la prédiction autrement dit la performance du modèle construit sur la base d'apprentissage. Dans ce cas, performant, signifie qu'il permet de classer tout individu avec le minimum d'erreurs possible. Cela présume l'existence d'un ensemble de test étiqueté pour pouvoir comparer les résultats obtenus par le système avec celles de l'expert. Dans cette section on va introduire les différentes Algorithmes utilisées pour comparer avec notre algorithme et les différentes mesures d'évaluations Avant de passer par la phase d'expérimentation et discussion des résultats.

Classfieur +	Expert		Total
	Positive	Négative	
Positive	TP	FP	TP+FP
Négative	FN	TN	FN+TN
Total	TP+FN	FP+TN	N

Tableau 4.3: Matrice de confusion

4.4.1 Mesures d'évaluation

L'efficacité de l'algorithme proposé a été validée en utilisant une gamme de base de traces de l'Université de Californie à Irvine (UCI) et KEEL (Derrac et al., 2015). Par conséquent, notre algorithme a été évalué en termes de précision de la classification et le taux de réduction des dimensions afin de montrer son comportement face à différents types de bases de traces. Nous devons considérer que notre algorithme est de sélection de pertinence et de classification. Par conséquent, la performance de notre algorithme doit être étudiée par ses performances en termes de précision de classification (Mokeddem et al., 2016).

Afin de valider correctement la procédure de classification, nous utilisons des mesures de performances sur les résultats de la classification. L'efficacité peut se définir selon plusieurs critères. Toutes les mesures courantes se basent sur la table de confusion dont un exemple est donné dans la table 4.3. Nous allons donner une définition formelle de ces mesures. Mais tout d'abord nous définissons les quatre notions suivantes pour une classe C_i:

- TP c'est l'ensemble de traces de la classe C_i bien classés ;
- FP est l'ensemble de traces assignés par erreur à la classe C_i ;
- FN est l'ensemble des traces de la classe C_i non classés C_i par le classificateur ;
- TN est l'ensemble des traces n'appartenant pas à la classe C_i et classé comme tels

4.4.1.1 Exactitude de classification

L'exactitude de classification est généralement la mesure la plus utilisée pour évaluer la performance d'une méthode en fouille de traces; elle est utilisé pour estimer combien la classification était correcte et performante. Afin de calculer ces métriques, nous calculons des termes comme, vrai positif (TP), vrai négatif (TN), faux négatifs (FN) et de faux positifs (FP) (tableau 4.3). Ensuite, la formule de calcul d'exactitude est donnée en (4.1). Inversement, il est connu que l'exactitude de la classification dépend de la méthode de sélection de traces pertinentes et la méthode de classification adoptée. Par conséquent, pour montrer la performance de l'algorithme proposé, nous avons choisi un ensemble de traces avec un certain nombre de méthodes de classification (Voir Section 2.3.2).

L'exactitude notée $\varepsilon(C_i)$ est la capacité du classificateur à bien classer les éléments qui lui sont soumis. C'est la somme du nombre de traces attribués à chaque classe par le système sur le nombre de traces que l'expert a attribué à chaque classe. Elle s'exprime de la façon suivante :

$$\varepsilon(C_i) = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.1}$$

4.4.1.2 Mesure de Kappa

En revanche, pour évaluer le comportement de l'algorithme proposé avec une base de traces déséquilibrées, nous avons utilisé la mesure Kappa de Cohen (Viera et al., 2005). La mesure Kappa est la plus couramment mesure utilisé qui produit la parfaite ampleur de l'accord entre observateurs (dans notre thèse nous avons deux observateurs le résultat de classificateur et l'expert). Kappa statistique est la variation entre les résultats expert / classificateur peut être mesurée en utilisant le Kappa, qui nous donne une évaluation chiffrée du degré d'accord. La mesure Kappa est la mesure la plus appropriée pour évaluer les résultats de classification des traces déséquilibrées. Par conséquent, elle regroupe l'exactitude de la classification pour chaque classe individuellement. Le calcul est basé sur la différence entre la quantité accord présente (classificateur) et combien l'accord devrait être présent (expert) (voir tableau 4.3). L'accord total est le pourcentage pour lequel les deux classificateur et l'expert est en accord : $P_o = TP+TN/N$ ou P_o est l'accord total. Kappa est une mesure de la différence entre P_o et P_e (voir 4.3), ou P_e est l'accord attendu (4.2):

$$P_e = \left(\frac{TP + FN}{N} \times \frac{TP + FP}{N}\right) \times \left(\frac{FP + TN}{N} \times \frac{FN + TN}{N}\right) \tag{4.2}$$

$$Kappa = \frac{P_o - P_e}{1 - P_e} \tag{4.3}$$

Les valeurs de la mesure Kappa varient dans une échelle de -1 à 1, où 1 est parfait accord, 0 est exactement un accord par chance, et les valeurs négatives indiquent accord moins que chance.

4.4.1.3 Mesure de Sensibilité

C'est une mesure qui indique la capacité du classificateur à classer correctement l'intégralité des traces. Formellement, elle s'exprime de la façon suivante :

$$\text{Sensibilité} = \frac{TP}{TP + FN} \tag{4.4}$$

La sensibilité permet de savoir si le classificateur est performant dans sa capacité à extraire de l'ensemble des traces ceux qui sont attribués à la classe en cours d'analyse tout en ayant peu d'oublis.

4.4.1.4 Mesure de Spécificité

La spécificité mesure comment la proportion des traces des patients qui ne présentent pas une maladie CAD peut être correctement exclu. Formellement la précision s'exprime de la façon suivante:

$$\text{Spécificité} = \frac{TN}{TN + FP} \tag{4.5}$$

Ce ratio permet de savoir en particulier si le classificateur, quand il classifie des une trace, n'affecte pas trop de traces à une classe par erreur.

4.5 Expérimentations et discussions

En conséquence, les expérimentations ont été effectuées sur deux groupes de bases de traces: dix bases de traces ont été sélectionnées à partir du référentiel de l'UCI (Asuncion and Newman, 2007) et à partir du référentiel KEEL (Derrac et al., 2015) nous avons choisi 24 bases de traces. Comme on peut le voir dans la table 4.2, ces bases de traces sont caractérisées par un grand nombre de descripteurs de traces et / ou un grand nombre de traces.

Tout d'abord, du référentielle UCI, on a sélectionné certaines des plus grands bases de traces de différents domaines (santé, gène, internet, champignons, et ondelettes). Deuxièmement, du référentielle KEEL un grand en ensemble de bases de traces de déséquilibrées de différents domaines d'application (de prédiction du cancer à partir de données de spectrométrie de masse, la reconnaissance de chiffres manuscrits, la classification de texte, la prévision de l'activité moléculaire et une data set artificielle) pour bien évaluer notre algorithme de sélection a été choisi. Les descripteurs sont de nature hétérogène: continues ou binaire, rares ou ondelettes, et toutes les bases de traces sont des problèmes de classification multiclasse.

Enfin, nous avons sélectionné des techniques FS / classification largement utilisés pour nos expériences SFS, C4.5, SVM, MLP et NB. Ces techniques sont décrites ci-dessus. Les expériences ont été effectuées en utilisant plateforme WEKA environnement (Hall et al., 2009). La validation a été faite utilisant la technique de validation croisée en calculant la moyenne (1 x 10 CV).

L'objectif de cette section est d'évaluer et de discuter les résultats obtenu avec notre algorithme proposé avec d'autres techniques de domaine. En raison de la haute dimensionnalité des traces, nous avons limité nos expériences avec l'algorithme SFS couplé avec les autres méthodes de fouille de traces (Mokeddem S et al., 2015). Comme cela a été décrit, les résultats expérimentaux sont structurés autour des deux principales questions suivantes:

1. L'impact de la réduction de l'espace des descripteurs sur la performance de la classification;
2. Le taux de de réduction de l'espace donnée par les approches.

Dans les sections suivantes, une analyse et une discussion des résultats est présenté. Tout d'abord,

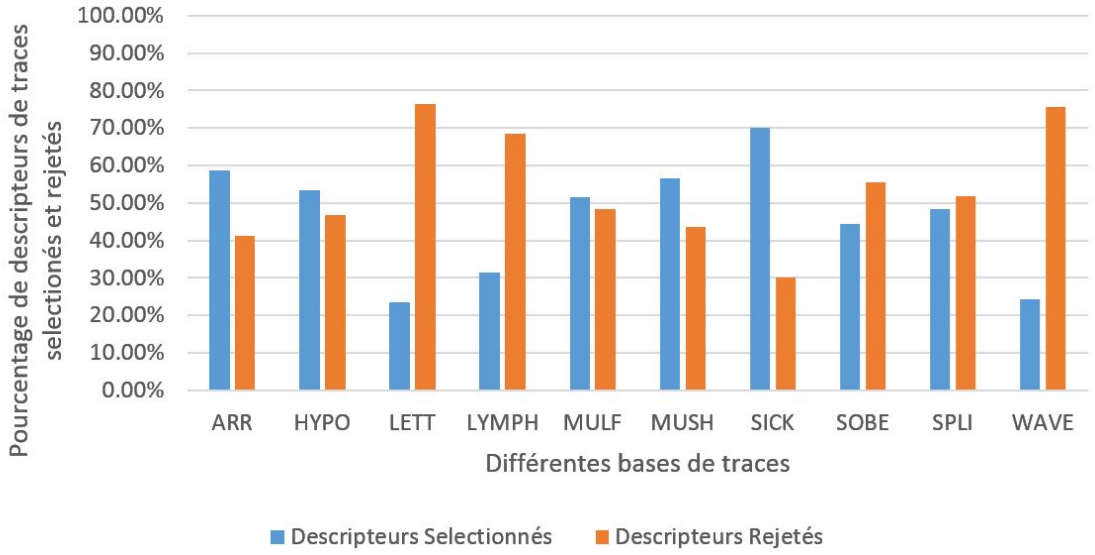


Figure 4.2: Taux de Sélection de descripteurs pertinents pour les différentes bases de traces du (UCI).

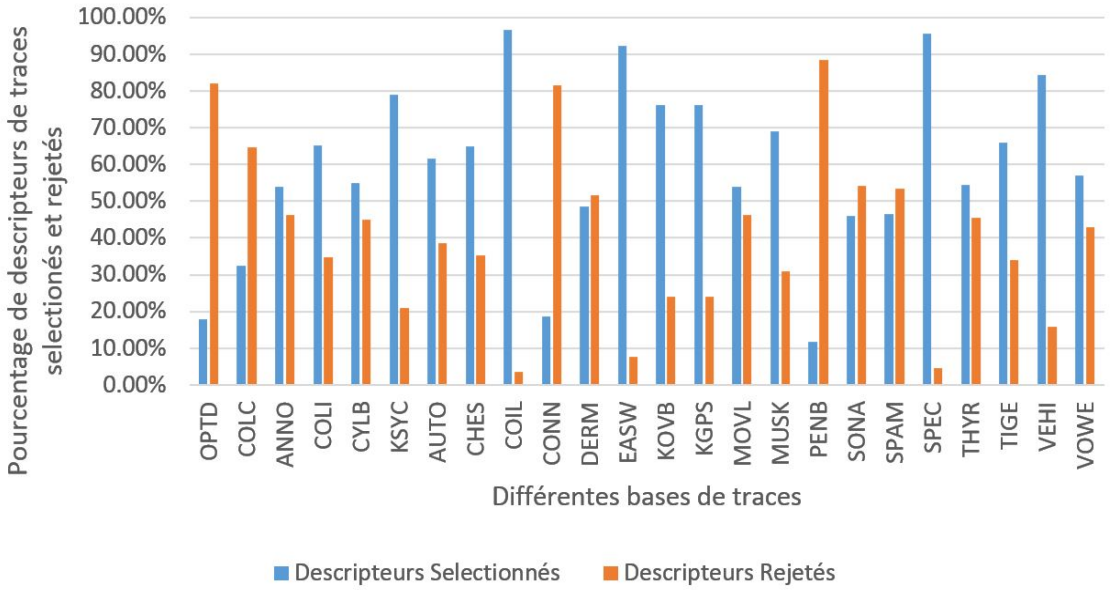


Figure 4.3: Taux de Sélection de descripteurs pertinents pour les différentes bases de traces (KEEL).

UCI Data sets	Algorithme Proposé	SFS		Première Approche Hybride (Ruiz et al, 2012)		Deuxième Approche Hybride (Ruiz et al, 2012)		CBFS	Original
		NB	C4.5	NB	C4.5	NB	C4.5		
ARRH	58.7	3.0	3.1	5.5	2.4	2.3	2.0	56.2	279
HYPO	53.3	23.3	20.3	15.9	14.5	10.3	10.7	53.3	29
LETT	23.5	72.5	63.1	68.8	68.8	39.4	45.0	25.0	16
LYMPH	31.5	26.3	N/A	N/A	N/A	N/A	N/A	16.6	18
MULF	51.6	2.4	2.1	3.4	3.2	2.1	1.5	93.8	649
MUSH	56.5	13.6	22.3	9.5	18.6	7.3	9.1	38.9	22
SICK	70.0	3.4	19.0	8.3	20.3	3.4	7.2	68.7	29
SOBE	44.4	55.5	N/A	N/A	N/A	N/A	N/A	50.0	36
SPLI	48.3	24.7	19.8	21.8	16.3	15.3	12.2	18.5	60
WAVE	24.4	32.3	18.2	30.5	24.0	21.3	17.5	57.5	40

Tableau 4.4: Comparaison du taux de Sélection des descripteurs pertinents (UCI)

nous commençons à discuter les résultats de l'exactitude de classification après nous allons discuter les résultats obtenus en utilisant la mesure Kappa. Dans ces expériences, le taux moyen des descripteurs de traces sélectionnées et les instances rejetées de l'ensemble d'apprentissage en fonction de la taille de la base de traces est présenté dans les (Figure 4.2, Figure 4.3, Table 4.4 et Table 4.5). Nous constatons que l'algorithme proposé pour la sélection des attributs de traces pertinents a une grande capacité de réduction. Il réduit la dimensionnalité jusqu'à 80%.

Malgré cette réduction drastique dans la dimensionnalités des traces d'apprentissage, nous allons montrer dans les expériences suivantes que ce processus ne dégrade pas les performances de classification mais au contraire une amélioration est constatée par rapport aux différentes méthodes de réduction. L'expérience sont approuvés, d'abord, nous avons calculé les exactitudes de notre algorithme couplé avec les différentes méthodes de fouille de traces (section 2.3.2). Les résultats des expériences sont évalués en utilisant l'exactitude afin d'afficher la pertinence de sous-ensemble de descripteurs pertinents produit par les différentes méthodes. Nous avons inclus deux autres techniques FS, une première technique hybride est la techniques de recherche SFS couplés avec les méthodes de fouille de traces (section 3.2.3). La seconde utilise les deux approches hybrides proposées dans (Ruiz et al., 2012). Les tables 4.6 et 4.7 montrent les exactitudes obtenues. Les quatre premières colonnes montrent les précisions obtenues en utilisant notre algorithme.

Les deux colonnes suivantes représentent les résultats obtenus avec l'aide de SFS Hybridé avec les évaluateurs NB et C4.5, respectivement. Les quatre colonnes suivantes correspondent aux deux approches hybrides proposées (Ruiz et al., 2012). Par lignes, nous trouvons la liste bases de traces utilisés dans les expériences. Basé sur les résultats de l'exactitude, nous avons accompli les comparaisons suivantes: il existe des différences significatives entre les exactitudes obtenues par l'algorithme proposé et les deux approches hybrides dans la plupart des bases de traces de la base UCI. Si nous analysons les résultats pour chaque base de traces, l'algorithme proposé à une meilleure exactitude de la classification que les autres techniques, sauf dans deux bases de traces, ce qui montrent l'efficacité de l'algorithme à choisir les descripteurs de traces les plus pertinents

KEEL data set	Algorithme proposé	ReliefF	CBF	INTERACT	CBFS	mRMR
OPTD	18.0	1.5	40.0	36.9	80.0	65
COLC	32.36	45.5	12.8	99.3	12.0	4.20
ANNO	53.8	48.7	30.7	82.0	51.2	39
COLI	65.2	4.3	73.9	73.9	48.8	23
CYLB	55.0	17.5	82.5	82.5	92.5	40
KSYC	79.0	0.0	22.5	3.2	70.9	62
AUTO	61.5	7.6	73.0	65.3	73.0	26
CHES	64.8	13.5	2.7	0.0	58.7	37
COIL	96.5	27.9	62.7	43.2	88.0	86
CONN	18.6	0.0	18.6	6.9	18.6	43
DERM	48.5	0.0	62.8	20.0	54.2	35
EASW	92.3	11.5	92.3	92.3	92.3	26
KOVB	76.1	35.7	95.2	95.2	76.1	42
KGPS	76.1	30.9	95.2	95.2	76.1	42
MOVL	53.8	0.0	36.2	25.2	0.46	91
MUSK	69.0	0.5	98.8	98.8	81.5	168
PENB	11.7	0.0	20.0	0.0	11.7	17
SONA	45.9	1.6	31.1	62.2	68.8	61
SPAM	46.5	0.0	24.1	1.7	42.0	58
SPEC	95.5	6.6	57.7	33.3	21.0	45
THYR	54.5	9.0	81.1	72.2	77.2	22
TIGE	65.9	61.2	99.1	99.1	93.1	232
VEHI	84.2	0.0	15.7	10.5	63.1	19
VOWE	57.1	7.1	26.3	40.0	57.1	14

Tableau 4.5: Comparaison du taux de Sélection des descripteurs pertinents (KEEL)

et à réduire la dimensionnalité des traces en optimisant le taux de classification.

Deux bases de traces bien connus ont aussi été utilisées: OPTD et COLC. Le but d'une méthode de FS est de sélectionner les attributs les plus pertinents et d'éliminer les non pertinents, ainsi que la détection les corrélé et les non-corrélés.

Table 4.7 présente les résultats obtenus Avec les différentes bases de traces des Référentielles UCI et KEEL parmi on y trouve: OPTD et COLC, respectivement. Dans COLC, notre algorithme a été en mesure de choisir le meilleur ensemble de caractéristiques, qui a conduit à 96,77% d'exactitude de classification obtenue par NB. En ce qui concerne, OPTD, il est intéressant aussi que la meilleure exactitude de classification a été également obtenue par notre algorithme (une exactitude de 98,79%).

La plupart des méthodes de classification FS améliorent la précision (tables 4.4, 4.5, 4.6, 4.7), dans certains cas. C'est parce que la plupart des descripteurs de traces (par exemple la taille de gènes mesurées dans une expérience de puces à ADN) ne sont pas pertinents pour une distinction précise entre les différentes valeurs de la classe, et par conséquent FS joue un rôle essentiel dans l'analyse des puces à ADN. Notre algorithme de FS maintien ou améliore la performance de classification en réduisant le nombre de descripteurs et en sélectionnant les attributs nécessaires. Dans certains ensembles de traces, le nombre d'attributs n'est pas si élevé, par conséquent, une grande amélioration n'était pas possible. Même alors, le meilleur résultat a été obtenu après l'application de l'algorithme proposé avec IB1. Bien que le résultat n'était pas très élevé avec ceux obtenus sans FS. Après l'utilisation de notre algorithme le résultat est un peu similaire, mais en diminuant sensiblement le nombre d'attributs (Tables 4.4, 4.4).

UCI Data sets	Algorithme Proposé					SFS		Première Approche Hybride (Ruiz et al, 2012)		Deuxième Approche Hybride (Ruiz et al, 2012)		CBFS				Original
	NB	IB1	C4.5	MLP	SVM	NB	C4.5	NB	C4.5	NB	C4.5	NB	IB1	SVM	C4.5	NB
ARRH	70.58	59.29	64.82	65.70	69.91	67.70	67.39	68.01	68.01	68.94	67.92	69.91	64.82	68.14	64.29	60.13
HYPO	95.22	86.69	99.57	96.81	93.61	94.96	99.30	95.10	99.07	94.92	98.90	94.64	93.10	93.13	99.36	95.32
LETT	66.05	96.32	88.32	79.45	78.45	65.67	85.17	65.67	84.99	55.74	80.50	64.11	96.00	82.34	84.45	63.97
LYMPH	87.83	81.08	79.72	80.40	87.83	82.43	82.43	n/a	n/a	n/a	n/a	78.37	83.78	85.81	77.02	83.10
MULF	94.40	96.25	88.05	98.00	97.60	96.87	93.11	97.21	92.42	96.80	93.74	95.00	96.25	97.55	92.74	94.37
MUSH	98.97	99.68	99.85	99.85	99.75	99.01	100	98.78	99.91	98.68	99.41	98.52	93.89	99.01	100	95.10
SICK	97.16	96.47	97.53	96.84	93.87	93.88	98.19	94.55	98.28	93.88	96.33	96.52	93.84	95.99	98.42	92.41
SOBE	93.41	88.14	90.92	92.83	94.43	92.67	92.97	n/a	n/a	n/a	n/a	89.31	98.31	93.85	91.50	92.97
SPLI	69.11	78.11	94.07	95.70	94.82	94.91	93.01	94.85	93.05	94.65	92.73	95.14	82.31	94.95	92.92	95.26
WAVE	82.05	71.94	82.04	81.86	85.22	81.55	75.44	80.85	76.20	80.38	75.93	80.12	79.20	87.02	74.75	80.02

Tableau 4.6: Résultats d'exactitude de différentes bases de traces (UCI)

KEEL Data sets	Algorithme Proposé				ReliffF		CBF			INTERACT		CBFS			mRMR	
	NB	IB1	C4.5	SVM	NB	C4.5	NB	C4.5	IB1	NB	C4.5	NB	IB1	SVM	C4.5	NB
OPTD	92.34	98.79	91.17	97.74	91.28	90.81	80.69	81.87	84.77	90.98	90.59	91.51	98.68	98.02	91.41	90.82
COLC	96.77	85.48	82.25	87.01	85.48	82.26	85.48	85.48	88.71	87.10	90.32	85.48	93.87	85.48	67.74	50.00
ANNO	90.08	91.75	87.97	84.96	75.27	90.86	75.27	90.98	95.99	63.02	90.80	63.02	90.80	79.95	90.98	75.27
COLI	85.86	81.25	85.59	82.88	78.53	85.32	81.53	85.59	81.79	83.15	81.79	83.15	81.79	81.79	85.58	80.70
CYLB	77.40	74.07	57.77	79.81	72.22	57.77	65.37	57.77	64.44	68.14	56.66	68.14	56.66	75.74	56.66	68.14
KSYC	97.66	95.16	90.33	97.83	94.66	91.83	93.14	92.00	94.33	94.83	92.50	95.00	91.50	97.00	82.00	79.32
AUTO	74.21	89.30	74.21	69.18	61.00	78.61	72.95	76.10	88.67	69.81	80.50	65.40	90.56	57.23	84.27	67.92
CHES	94.52	93.02	97.43	94.18	88.39	94.93	87.82	94.05	89.98	87.82	94.05	91.99	92.24	93.92	94.05	91.99
COIL	94.03	90.85	94.03	94.03	78.38	93.99	84.64	93.96	90.03	79.57	93.95	92.49	89.55	94.06	94.03	93.40
CONN	72.69	68.33	78.28	n/a	72.14	80.96	72.20	81.01	67.03	79.57	80.76	70.00	57.50	n/a	70.44	70.00
DERM	98.90	95.90	95.62	96.99	97.54	96.44	96.99	96.72	94.80	97.54	94.80	96.44	94.26	96.99	94.80	97.54
EASW	100	100	100	100	94.83	100	100	100	100	100	100	100	100	100	100	100
KOVB	100	99.95	99.95	99.86	100	99.95	99.95	99.95	99.95	100	99.95	99.91	100	100	99.95	96.00
KGPS	100	100	99.93	99.93	100	99.87	97.74	99.81	99.93	100	100	100	99.93	99.87	99.87	100
MOVL	64.16	86.11	68.61	70.55	62.77	67.77	60.00	66.94	86.00	63.61	70.27	64.72	84.72	67.50	69.16	58.88
MUSK	85.92	97.05	99.15	99.57	75.84	99.15	99.15	99.15	98.31	99.15	99.15	83.40	98.52	99.15	97.89	82.56
PENB	85.72	97.18	87.90	95.27	84.63	88.81	84.00	87.63	96.63	83.27	89.63	83.54	97.03	94.72	89.18	83.81
SONA	75.48	87.01	75.96	79.32	67.78	72.11	67.30	76.92	85.57	67.78	79.80	69.71	85.57	78.36	78.84	69.26
SPAM	90.42	89.90	92.29	88.27	79.68	92.93	77.37	92.60	90.68	79.74	92.93	78.96	90.95	86.68	93.15	89.31
SPEC	79.40	67.04	79.40	79.40	69.28	76.40	73.40	77.52	74.26	70.78	76.40	70.78	73.78	77.90	76.77	72.28
THYR	95.41	93.47	98.61	92.63	95.00	98.61	94.58	98.61	97.91	95.27	98.61	94.58	96.94	92.62	98.50	94.58
TIGE	87.86	99.75	99.75	99.91	75.32	98.93	99.91	99.91	99.91	99.91	99.91	77.54	99.91	99.91	99.91	94.75
VEHI	77.54	67.13	99.15	74.94	69.14	77.89	68.79	79.31	75.76	69.14	78.60	72.10	73.64	74.94	70.27	63.61
VOWE	96.65	99.89	99.39	96.45	93.62	89.88	93.52	98.17	99.89	93.21	89.17	93.92	99.49	94.73	98.07	93.92

Tableau 4.7: Résultats d'exactitude de différentes bases de traces (KEEL)

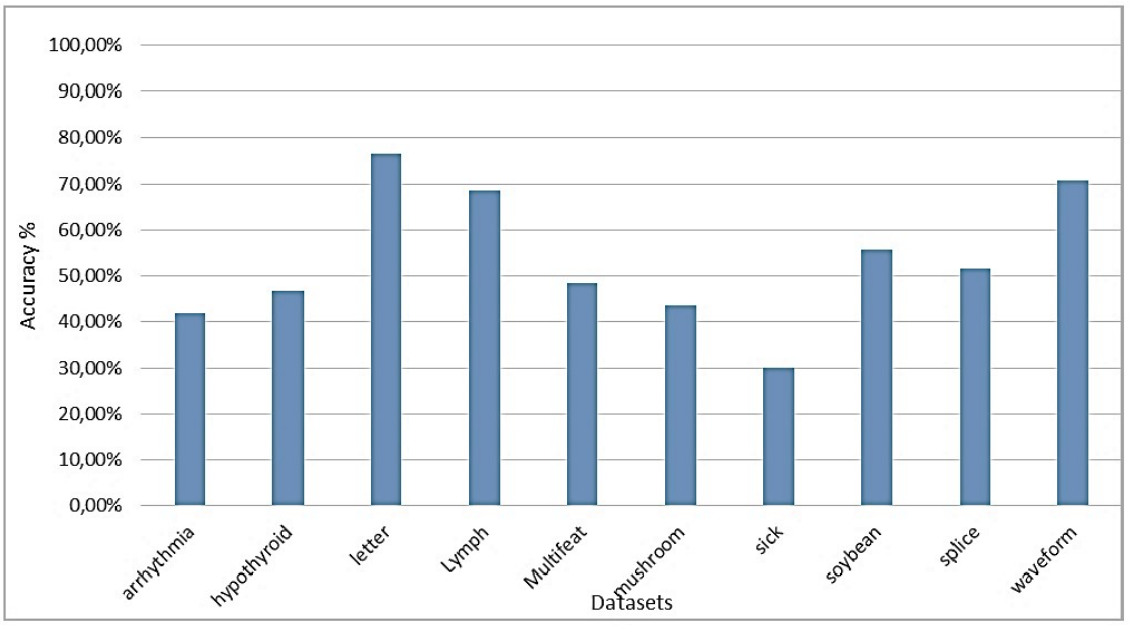


Figure 4.4: Exactitude de notre algorithme utilisant le référentiel UCI data sets.

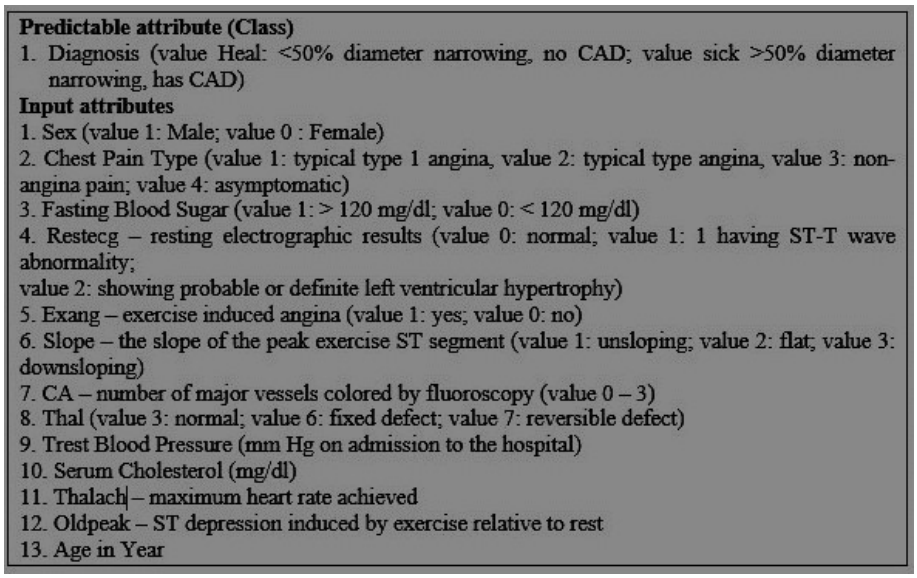
Le coefficient Kappa est utilisé pour la comparaison. La mesure Kappa tient compte de tous les éléments de la matrice de confusion (voir Section 4.4.1.2). Les tables 4.8 et 4.9 présentent la Kappa obtenu. Trois méthodes FS ont été utilisées et évalués individuellement en utilisant quatre méthodes wrapper. Pour la première base de traces, le BFS fournit le meilleur Kappa 0,9026, avec une amélioration significative de l'aide IB1. Pour le reste des ensembles de traces notre algorithme offre et valorise la meilleur Kappa 0.6921, 0.5781, 0.8347 et 0.978 pour colic, cylinder-bands, kdd_synthetic_control et page-blocks respectivement. Les résultats présentés dans tables 4.8 et 4.9 montrent une efficacité claire de notre technique de FS et ses performances. Selon cette constatation, les descripteurs obtenus par l'algorithme proposé sont ceux utilisés par des experts. Ceci s'explique par le fait que la plupart des valeurs Kappa sont plus grandes que 0.5, ce qui est considérable.

UCI Data sets	Algorithme Proposé				SFS		Premiere Approche Hybride (Ruiz et al, 2012)		Deuxieme Approche Hybride (Ruiz et al, 2012)		CBFS			Original	
	NB	IB1	C4.5	MLP	SVM	NB	C4.5	NB	C4.5	NB	C4.5	NB	IB1	SVM	C4.5	NB
ARRH	0.54	0.26	0.45	0.46	0.48	0.50	0.52	n/a	n/a	n/a	n/a	0.53	0.44	0.43	0.46	0.44
HYPO	0.59	0.32	0.97	0.76	0.29	0.97	0.57	n/a	n/a	n/a	n/a	0.52	0.57	0.20	0.97	0.60
LETT	0.64	0.96	0.87	0.78	0.77	0.64	0.87	n/a	n/a	n/a	n/a	0.62	0.95	0.81	0.87	0.62
LYMPH	0.77	0.63	0.61	0.62	0.76	0.66	0.55	n/a	n/a	n/a	n/a	0.58	0.69	0.72	0.57	0.67
MULF	0.93	0.95	0.86	0.97	0.97	0.90	0.87	n/a	n/a	n/a	n/a	0.94	0.95	0.97	0.87	0.91
MUSH	0.97	0.99	0.99	0.99	0.99	0.99	0.99	n/a	n/a	n/a	n/a	0.97	0.87	0.98	1	0.91
SICK	0.74	0.68	0.78	0.71	0.06	0.60	0.80	n/a	n/a	n/a	n/a	0.68	-0.05	0.63	0.89	0.52
SOBE	0.92	0.86	0.90	0.91	0.93	0.91	0.91	n/a	n/a	n/a	n/a	0.88	0.88	0.93	0.90	0.92
SPLI	0.93	0.66	0.90	0.93	0.91	0.92	0.89	n/a	n/a	n/a	n/a	0.93	0.72	0.93	0.90	0.92
WAVE	0.73	0.57	0.62	0.72	0.77	0.73	0.64	n/a	n/a	n/a	n/a	0.70	0.68	0.80	0.62	0.70

Tableau 4.8: Résultats de KAPPA de différentes bases de traces (base UCI)

KEEL Data sets	Algorithme Proposé				ReliefF		CBF			INTERACT		CBFS			mRMR	
	NB	IB1	C4.5	SVM	NB	C4.5	NB	C4.5	IB1	NB	C4.5	NB	IB1	SVM	C4.5	NB
OPTD	0.91	0.98	0.90	0.97	0.89	0.90	0.83	0.86	0.91	0.90	0.89	0.90	0.89	0.98	0.87	0.84
COLC	0.86	0.56	0.37	0.78	0.71	0.75	0.63	0.86	0.70	0.77	0.65	0.71	0.37	0.55	0.73	0.60
ANNO	0.77	0.77	0.63	0.60	0.55	0.76	0.55	0.76	0.89	0.34	0.76	0.34	0.76	0.33	0.76	0.55
COLI	0.69	0.60	0.67	0.62	0.54	0.67	0.60	0.67	0.60	0.64	0.60	0.64	0.60	0.61	0.67	0.59
CYLB	0.53	0.44	0.48	0.57	0.42	0.48	0.22	0.48	0.23	0.30	-0.04	0.49	-0.01	0.49	-0.01	0.30
KSYC	0.97	0.94	0.88	0.97	0.93	0.90	0.91	0.90	0.93	0.93	0.91	0.94	0.89	0.96	0.78	0.75
AUTO	0.66	0.86	0.66	0.59	0.49	0.72	0.64	0.68	0.85	0.60	0.74	0.54	0.87	0.41	0.79	0.57
CHES	0.89	0.86	0.94	0.88	0.76	0.89	0.75	0.88	0.79	0.75	0.88	0.83	0.84	0.87	0.88	0.83
COIL	0.11	0.01	0.11	0.05	0.10	0	0.14	0	0.09	0.12	0	0.07	0.06	0.11	0.11	0.06
CONN	0.32	0.36	0.52	n/a	0.33	0.59	0.32	0.59	0.33	0.12	0.59	0.23	0.16	n/a	0.28	0.23
DERM	0.98	0.94	0.94	0.96	0.96	0.95	0.96	0.95	0.93	0.96	0.93	0.95	0.92	0.96	0.93	0.96
EASW	1	1	1	1	0.89	1	1	1	1	1	1	1	1	1	1	1
KOVB	1	0.98	0.98	0.94	1	0.98	0.98	0.98	0.98	1	0.98	0.96	1	1	0.98	0.96
KGPS	0.61	0.85	0.66	0.68	0.60	0.65	0.57	0.64	0.84	0.61	0.68	0.62	0.83	0.65	0.66	0.55
MOVL	0.71	0.94	0.98	0.99	0.51	0.98	0.98	0.98	0.96	0.98	0.98	0.66	0.97	0.98	0.95	0.65
MUSK	0.84	0.96	0.86	0.94	0.82	0.87	0.82	0.86	0.96	0.81	0.88	0.81	0.96	0.94	0.87	0.82
PENB	0.84	0.96	0.86	0.94	0.82	0.87	0.82	0.86	0.96	0.81	0.88	0.81	0.96	0.94	0.87	0.82
SONA	0.51	0.73	0.51	0.58	0.36	0.44	0.35	0.53	0.70	0.36	0.59	0.40	0.70	0.56	0.57	0.37
SPAM	0.79	0.78	0.83	0.74	0.60	0.85	0.56	0.84	0.80	0.60	0.85	0.59	0.81	0.71	0.85	0.78
SPEC	0.34	0.01	0.32	0.20	0.36	0.27	0.43	0.31	0.24	0.39	0.26	0.39	0.23	0.32	0.28	0.41
THYR	0.57	0.44	0.90	0.03	0.53	0.90	0.47	0.90	0.85	0.55	0.90	0.47	0.76	0.03	0.90	0.47
TIGE	0.75	0.99	0.99	0.99	0.49	0.97	0.99	0.99	0.99	0.99	0.99	0.53	0.99	0.99	0.99	0.89
VEHI	0.21	0.14	0.99	0.25	0.28	0.39	0.27	0.40	0.32	0.27	0.38	0.29	0.30	0.25	0.68	0.61
VOWE	0.77	0.99	0.96	0.74	0.68	0.93	0.68	0.89	0.99	0.66	0.81	0.67	0.96	0.57	0.88	0.67

Tableau 4.9: Résultats de KAPPA de différentes bases de traces (base KEEL)


Predictable attribute (Class)
1. Diagnosis (value Heal: <50% diameter narrowing, no CAD; value sick >50% diameter narrowing, has CAD)

Input attributes
1. Sex (value 1: Male; value 0 : Female)
2. Chest Pain Type (value 1: typical type 1 angina, value 2: typical type angina, value 3: non-angina pain; value 4: asymptomatic)
3. Fasting Blood Sugar (value 1: > 120 mg/dl; value 0: < 120 mg/dl)
4. Restecg – resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality;
value 2: showing probable or definite left ventricular hypertrophy)
5. Exang – exercise induced angina (value 1: yes; value 0: no)
6. Slope – the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: downsloping)
7. CA – number of major vessels colored by fluoroscopy (value 0 – 3)
8. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
9. Trest Blood Pressure (mm Hg on admission to the hospital)
10. Serum Cholesterol (mg/dl)
11. Thalach – maximum heart rate achieved
12. Oldpeak – ST depression induced by exercise relative to rest
13. Age in Year

Figure 4.5: Liste des descripteurs des patients du CAD.

4.6 Évaluation expérimentales de la pertinence des traces des patients pour les maladies CAD

Pour les expérimentations, les patients ont été évalués en utilisant 14 descripteurs de traces. L'ensemble des traces est prise à partir du référentiel UCI (Blake and Merz, 1998). Pour en finir avec notre système est validé en utilisant des traces des patients de l'hôpital de Cleveland. Des descripteurs telles que l'âge, le sexe, le type de douleur à la poitrine, au repos la pression artérielle, le cholestérol sérique en mg / dl, la glycémie à jeun, reposant sur les résultats de l'électrocardiographies, et la fréquence cardiaque maximale atteinte, l'exercice angor, dépression du segment ST, la pente de l'exercice pic du segment ST , nombre de gros vaisseaux, et le diagnostic CAD sont présentés. Figure 4.5.

Les expériences sont approuvés et pour chacune des méthodes wrappers proposées. On calcule d'abord l'exactitude moyenne des algorithmes wrapper couplés avec BN classificateur. Ainsi, nous faisons usage de 10-validation croisée, en outre, chaque stratégie de wrapper produit un sous-ensemble de descripteur et nous avons fait des expériences pour mesurer l'efficacité de ces sous-ensembles avec l'utilisation des méthodes décrites dans la section 3.2.3 utilisés pour le diagnostic CAD. Afin de montrer la pertinence des stratégies FS, examinable à partir de la table 4.11 qui comportent les résultats des algorithmes de sélection produit comparé avec l'ensemble de traces originales (sans FS). Par ailleurs, la table 4.10 fournit la liste des descripteurs du CAD et les sous-ensembles de traces sélectionné par les algorithmes de wrapper.

Comme il est illustré dans la table 4.11, la plus haute exactitude est obtenue par l'algorithme GA-BN. Cette exactitude est de 85,82% et c'est la meilleure valeur rapportés dans la littérature

N°	Descripteurs	Descripteurs sélectionnés					
		GA wrapper BN	GA wrapper SVM	GA wrapper MLP	GA wrapper C4.5	BFS wrapper BN	SFFS wrapper BN
1	Chest pain type: cp	√	√	√	√		√
2	Age		√	√			
3	Sex	√		√			
4	Resting blood pressure: restbps			√			
5	Cholesterol: chol					√	
6	Fasting blood sugar: fbs				√	√	
7	Resting electrocardiographic results: restescg	√					√
8	Maximum heart rate achieved: thalach					√	√
9	Exercise induced angina: exang		√			√	
10	ST depression induced by exercise relative to rest: oldpeak	√	√				√
11	The slope of the peak exercise ST segment: slope	√	√	√			
12	Number of major vessels(0-3) colored by fluoroscopy: ca	√	√	√	√	√	√
13	Thalium: thal	√	√	√	√	√	√

Tableau 4.10: Descripteurs sélectionnés par différentes méthodes Wrapper

Wrapper Algorithms	Accuracy of different ML methods			
	BN	SVM	MLP	C4.5
GA wrapper	85.82	83.82	79.86	78.54
BFS wrapper	83.5	80.53	80.53	78.55
SFFS wrapper	84.49	83.17	77.89	78.22
Without FS	82.5	83.17	79.2	76.57

Tableau 4.11: Évaluation de performances des différents Méthodes Wrappers

(figure 4.6). Par exemple, (Abidin et al., 2009; Anooj, 2012; Hedeshi and Abadeh, 2014; Setiawan et al., 2009; Tsipouras et al., 2008) avait atteint 57,58% précisions, 73,40, 79,75% et 85,72%, ce qui est inférieur à la valeur rapporté dans cette étude.

Le gain de temps de notre algorithme est très important et devient évident lorsque la taille des descripteurs est diminuée. Cette analyse comparative indique que l'algorithme proposé est une très bonne technique de FS qui offrent la meilleure performance de classification. GA-BN a pu choisir le meilleur ensemble de d'attributs, qui a conduit à 85,82% de précision de classification obtenu (Figure 4.6).

4.7 Conclusion et perspectives

FS a été un champ actif et productif de la recherche en fouille de traces ce qui fait preuve par son efficacité dans l'amélioration de la performance du système de classification et dans la prise de décision en diminuant la complexité du modèle. Toutefois, la sélection de la méthode FS convenable n'est pas tâche triviale. Dans ce chapitre, nous avons validés notre algorithme de FS qui est approprié pour traiter différents types de problèmes d'analyse de traces. Des expérimentations sur un ensemble de 14 FS méthodes appliquées sur 34 ensembles de données UCI /KEEL a été présenté visant à étudier leur performance et montrant la performance de notre algorithme de FS vis-à-vis les méthodes existantes dans la littérature.

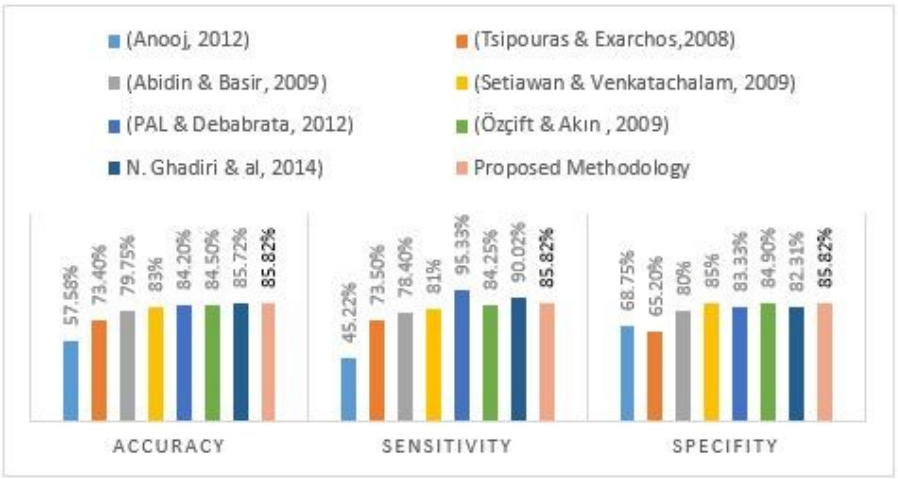


Figure 4.6: Analyse des différents systèmes.

Dans le domaine de la recherche de FS, trois approches principales ont été évalués: filtres, wrapper et les méthodes hybrides. Pour tester l'efficacité, quatre mesures d'évaluation a été employé pour mesurer l'efficacité de l'algorithme. Exactitude, Sensibilité et Spécifité de classification pour évaluer la précision des différentes méthodes et de la Kappa mesure le degré d'accord entre le classificateur et l'expert. Enfin, nous pouvons affirmer que notre algorithme de FS fournit des résultats prometteurs. En outre, la stratégie SFS n'est pas une bonne méthode pour les grands ensembles de traces en raison de sa lente convergence. Enfin, à partir de la comparaison des résultats, il est clair que notre proposition surclasse les résultats publiés dans la littérature. Cette amélioration est assez grande pour être scientifiquement et pratiquement intéressante. Notre algorithme peut être considéré comme un complément prometteur à des techniques existantes de fouille de traces.

Dans les études futures, l'algorithme proposé peut être exécuté en parallèle pour la génération des chromosomes et l'évaluation de chaque chromosome. Nous allons essayer d'améliorer constamment l'algorithme proposé pour différents problèmes de décision.

Chapitre 5

Un CDSS Basé Logique Floue pour l'Aide au Diagnostic des CADs

Sommaire

5.1 Introduction

Nous avons vu dans les premiers chapitres que la sélection des traces pertinentes a un impact important dans l'amélioration d'un système d'aide à la décision en générale. Les systèmes d'aide à la décision sont présents dans de nombreux domaines et ont pour objectif d'aider le décideur dans sa tâche en lui fournissant tous les éléments pertinents pour la prise de décision. Ces éléments pertinents sont généralement représentés sous forme de connaissances. En effet, ces connaissances dont disposent les humains sur le monde et qui peuvent alimentées un système d'aide à la décision ne sont presque jamais parfaites. Ces imperfections peuvent être distinguées en deux classes :

- Imprécisions pour désigner les connaissances qui ne sont pas aperçues ou définies clairement. Par exemple, au lieu de dire qu'une personne mesure 2 mètres et 10 centimètres, nous disons habituellement que cette personne est très grande.
- Incertitudes pour indiquer les connaissances dont la validité est en doute. Par exemple, si nous savons qu'une personne s'est cognée la tête sur un panneau de basket Ball, nous devinons qu'elle est probablement très grande.

De nos jours, la logique floue (fuzzy logic) est un axe de recherche important sur lequel de nombreux scientifiques travaillent. De Nombreux travaux et d'application sont déjà disponible dans différents domaines: le domaine grand public (appareils photos, machines à laver, fours à micro-onde), le domaine industriel (classification, aide à la décision industrielle, aide à la décision médicale, réglage et commande de processus, aux transports, à la transformation de la matière, à la robotique).

Les formalismes de la logique floue ont été formulés en 1965 par le professeur Lotfi A. Zadeh, de l'Université de Berkeley en Californie (Zadeh, 1965). Il a introduit la notion de sous-ensemble flou pour fournir un moyen de représentation et de manipulation des différents aspects de la connaissance humaine : l'incertitude et l'imprécision. Dès 1975, Mamdani et Assilian publient les premiers résultats exploitant cette théorie dans des systèmes de réglage (Mamdani and Assilian, 1975). En usant une structure de contrôleur relativement simple, ils ont obtenu de meilleurs résultats lors de la commande de certains processus que ceux fournis par un régulateur standard de type PID.

Dans le cadre des traces des patients du CAD, le prétraitement des traces est une procédure qui a comme objectif de préfiltrer les descripteurs représentant ces traces et qui ne sont pas informatifs (voir chapitres 3 et 4). Dans ce chapitre nous envisageons deux tâches celle de la normalisation des traces des patients du CAD et celle de la construction du CDSS à l'aide de la logique floue afin de faciliter et d'améliorer la performance.

Dans la section qui suit, nous décrivons les notions de base et les connaissances préliminaires de la logique floue et les Systèmes d'inférence Flous (SIF). Dans la section 3, nous détaillons les

composants principaux des SIF tels que la fuzzification, l'inférence et la défuzzification pour le traitement des traces.

Dans la section 4, nous expliquons l'approche floue que nous proposons pour la construction du CDSS. Cette approche est basée sur le système d'inférence proposé (Mamdani and Assilian, 1975), qui conduit à inférer une bases de connaissances floue. La défuzzification engendre la solution selon l'inférence et nous résulte la décision finale. Dans la section 5, nous proposons un protocole expérimental pour évaluer l'impact de notre modélisation flou dans le processus de la construction d'un CDSS pour le CAD.

5.2 Logique Floue

La logique classique ne comprend que des propositions soit avec une valeur soit vraie, soit fausse (1 ou 0). Par exemple, la logique classique peut facilement partitionner la fréquence cardiaque d'une personne en deux sous-ensembles, «moins de 70 battements par minute (bpm)» et «70 bpm ou plus». La figure 5.2 illustre le résultat de cette partition. Toutes les fréquences cardiaques de moins de 70 sont alors considérées comme appartenant à l'ensemble «moins de 70 bpm». On leur affecte une valeur de 1. Toutes les fréquences aboutissant plus que 70 bpm ne sont pas considérées comme appartenant à l'ensemble «moins ou égales à 70 bpm ». On leur attribue une valeur de 0.

Néanmoins, le raisonnement humain s'appuie souvent sur des connaissances ou des données inexactes, incertaines ou imprécises. Une personne avec la fréquence cardiaque soit de 69.05 bpm soit de 70.05 bpm, évidement ne fera pas de distinction entre ces deux valeurs. Cette personne sera pourtant capable de dire si la fréquence de battement de son cœur est «normal» ou «Haute», sans pour cela utiliser de la fréquence cardiaque limite ni de mesure précise.

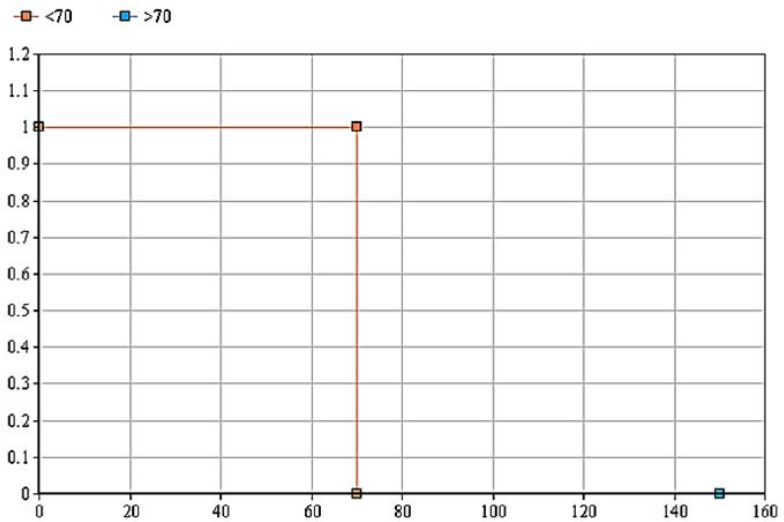


Figure 5.1: Sous ensemble de la fréquence cardiaque selon la logique classique

Malgré ce flou qui détermine notre perception du monde dont la qualité des raisonnements est exceptionnelle que ce soit la nature de la situation (simple ou complexe), les décisions prises sont en général très bonnes par rapport au manque de précision et à l'incertitude des données du problème. La logique floue est une généralisation de la logique classique, En introduisant la notion de degré dans la vérification d'une condition, permettant ainsi à une condition d'être dans un autre état que vrai ou faux, la logique floue attribue une extensibilité très perceptible aux raisonnements qui l'utilisent, ce qui rend possible la prise en compte des imprécisions et des incertitudes.

5.2.1 Sous-ensembles floue

Cependant en théorie des ensembles classiques, l'appartenance d'un élément à un sous-ensemble est booléenne. Les sous-ensembles flous permettent en revanche de connaître le degré d'appartenance d'un élément au sous-ensemble Un sous-ensemble flou A d'un univers du discours U est caractérisé par une fonction d'appartenance (Zadeh, 1965) :

$$\mu_A : U \longrightarrow [0, 1] \tag{5.1}$$

Où μ_A est le degré d'appartenance dans l'univers du discours U dans le sous-ensemble flou.
On peut définir aussi un sous-ensemble flou $\overline{A}$ comme suit (Zadeh, 1965) :

$$\overline{A} = \{(x, \mu_{\overline{A}}(x))/x \in U\} \tag{5.2}$$

Où $\mu_{\overline{A}}(x)$ est le degré d'appartenance de x dans $\overline{A}$

Example : Soit U défini sur $\Re$ et A le sous-ensemble classique pour représenter les fréquences cardiaque inférieurs ou égales à 70 bmp; alors, nous avons :

$$A = \{(x, \mu_A(x))/x \in U\}$$

Où la fonction caractéristique est définie par :

$$\mu_A(x) = \begin{cases} 1 & si\ x \leq 70, \\ 0 & si\ x > 70, \end{cases}$$

Qui est montré dans la figure 1. Soit, alors, un sous-ensemble flou $\overline{A}$ qui représente les nombres réels proches de 70, nous avons donc comme fonction caractéristiques:

$$\overline{A} = \{(x, \mu_{\overline{A}}(x))/x \in U\}$$

Où la fonction caractéristique est définie comme suit :

$$\mu_{\overline{A}}(x) = \frac{1}{1 + 10(x - 70)^2}$$

Et qui est montrée dans la figure 5.2 :

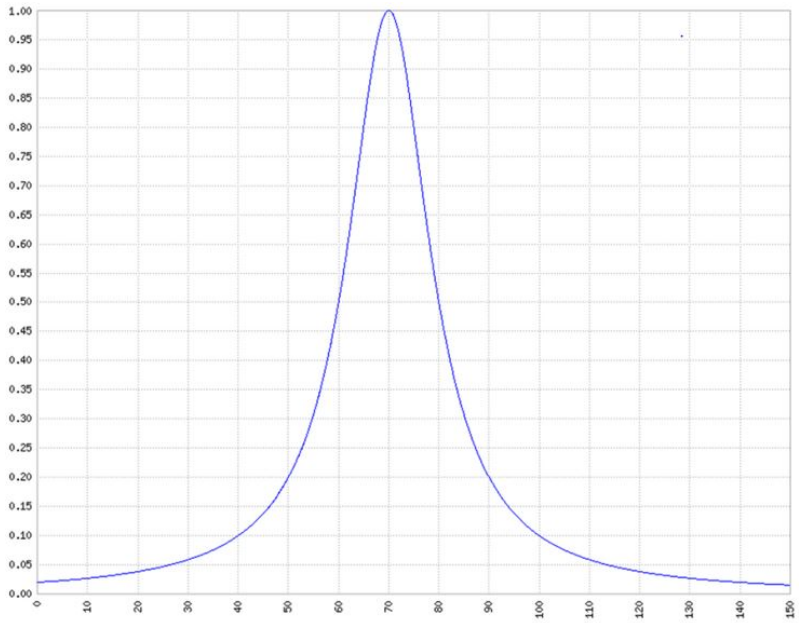


Figure 5.2: Sous ensemble de la fréquence cardiaque selon la logique Floue

5.2.2 Variables linguistiques

En logique floue les concepts des systèmes sont usuellement représentés par des variables linguistiques. Une variable linguistique permettra de définir des systèmes flous en langage naturel, la fonction d'appartenance établie le lien entre logique floue et variable linguistique. Une variable linguistique est définie par (Lin and Lee, 1991)

$$(X, U, T(X), \mu_x) \tag{5.3}$$

Où X désigne le nom de la variable, U est l'univers du discours associé à X, est $T(X) = \{T_1, T_2, T_3, ..., T_n\}$ l'ensemble des valeurs linguistiques de la variable X (appelé également termes linguistiques), et finalement μ_x sont les fonctions d'appartenance associées à l'ensemble de termes linguistiques. Exemple continuant dans l'exemple de la fréquence cardiaque, donc la variable linguistique associé à la variable Fréquence cardiaque est définit comme suit :

$$(\text{Fréquence cardiaque}, U = 0, 150, T(X) = \{T_{x1}, T_{x2}, T_{x3}\}, \mu = \{\mu_{x1}, \mu_{x2}, \mu_{x3}\}) \tag{5.4}$$

Où U est l'univers des tailles humaines défini en centimètres, l'ensemble T est constitué de trois valeurs linguistiques: T_{x1} = bradycardie, T_{x2} = normale et T_{x3} = tachycardie, et les fonctions d'appartenances définies par chaque terme linguistique sont:

$$\mu_{bradycardie} = trapzoide(x, \alpha, 30, 50, 200)$$

$$\mu_{bradycardie} = trapzoide(x, 70, 85, 95, 110)$$

$$\mu_{tachycardie} = trapzoide(x, 90, 150, 160, \beta)$$

Par entente nous notons désormais α et β comme les limites des trapézoïdes qui sont reportées au-delà de l'univers de discours. Cela veut dire que les valeurs en dehors de l'intervalle $]\alpha, \beta[$ ont toujours un degré d'appartenance de 1. La fonction d'appartenance triangulaire est définie comme suit (Jang et al., 1997)

$$triangulaire(x; a, b, c) = max\left(min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right). \tag{5.5}$$

Avec $(a < b < c)$ où b est le sommet du triangle tandis que a et c imposent la largeur du domaine de la valeur à fuzzifier. Nous montrons notre exemple dans la Figure 5.3 où on représente la variable linguistique fréquence cardiaque de l'être humain. La définition de chaque sous-ensemble flou repose sur l'intuition cardiologues. Si une personne a une fréquence cardiaque 70 bpm cela se traduira par différents degrés d'appartenances à chacun des sous-ensembles flous :

$$\mu_{bradycardie} = 0, \mu_{normale} = 1, \mu_{tachycardie} = 0$$

Nous pouvons conclure que ce patient appartient au sous-ensemble des patients qui ont une fréquence cardiaque normale Par contre un autre patient qui a une fréquence cardiaque de 100 bpm a les appartenances suivantes:

$$\mu_{bradycardie} = 0, \mu_{normale} = 0.4, \mu_{tachycardie} = 0.18$$

Ce patient peut donc être considéré à la fois de fréquence cardiaque normale et tachycardie avec une plus forte appartenance à la fréquence " normale ". Si nous voulons traiter cet exemple pour un domaine en particulier, la définition de sous-ensembles changera, par exemple pour la fréquence cardiaque en repos ou bien la fréquence cardiaque en activité.

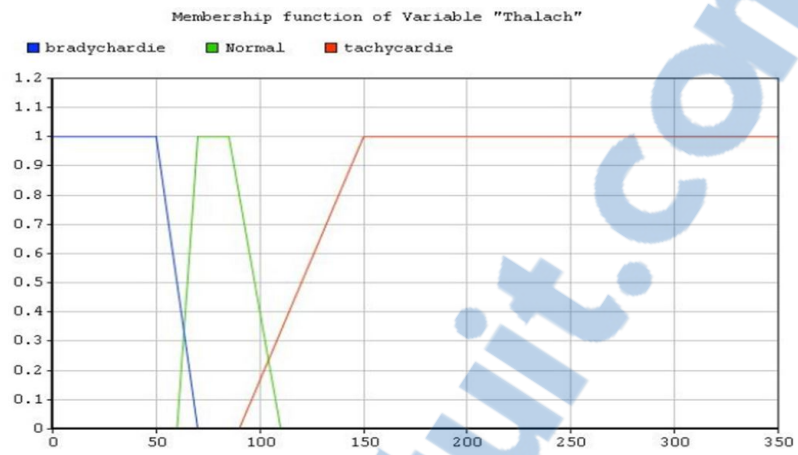


Figure 5.3: Fonctions d'appartenance de la variable fréquence cardiaque

5.2.3 Système d'Inférence Flou

Les systèmes d'inférence floue (SIF) sont une des applications les plus usuelles de la logique floue. Ils implémentent des concepts, sous la forme de variables linguistiques, ainsi qu'un raisonnement déductif, à l'aide de règles floues. Par conséquent, un SIF a comme but de transformer les données ordinaire d'entrées en données de sortie en les infèrent avec un ensemble des règles floue. Les entrées sont issues du processus de fuzzification et l'ensemble de règles généralement sont définies ou bien validées par le savoir-faire de l'expert voir Figure 5.4. Un SIF est composé de trois modules: a) Fuzzification, b) Inférence et c) Défuzzification. La première partie est la fuzzification, qui comporte les variables linguistiques utilisées dans le système. Il s'agit donc d'une transformation des entrées réelles en une partie floue définie sur un espace de représentation lié à l'entrée. Cet espace de représentation est évidemment un sous-ensemble flou. Durant l'étape de la fuzzification, chaque variable d'entrée et de sortie est liée à des sous-ensembles flous.

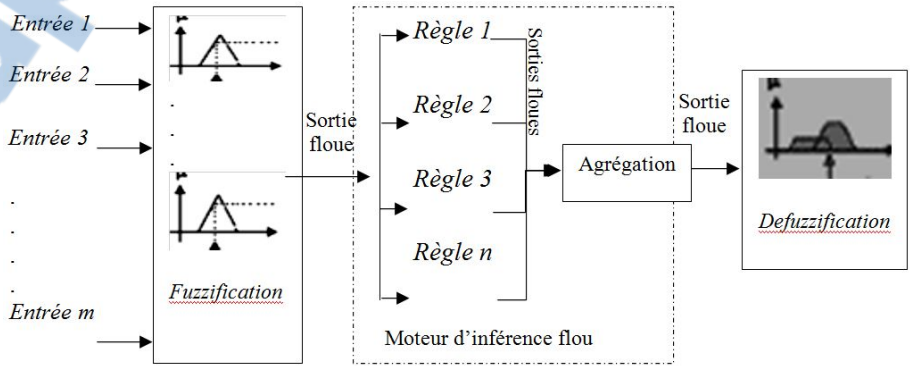


Figure 5.4: Structure d'un SIF

Le deuxième composant est le moteur d'inférence (MI), le MI contient généralement une base de règle floue les plus courantes, dites conjonctives. Chaque règle délivre une conclusion partielle qui est ensuite agrégée avec les sorties des différentes règles pour fournir une conclusion. Dans la suite de ce chapitre on va détailler les différentes étapes dans un cadre formel. Le troisième module est la défuzzification, ce module permet l'opération inverse de la fuzzification et permet de transformer les sorties floues de l'inférence en une valeur non floue comme réponse finale du SIF.

5.2.3.1 Fuzzification

Pour montrer le processus de fuzzification il faut d'abord savoir combien des variables d'entrée seront définies dans le SIF. Nous rappelons qu'une variable d'entrée (fréquence cardiaque, tension artérielle, taux de cholestérol, poids, taille, Age, etc.) est un paramètre qui prend ses valeurs dans un univers de discours de $\Re$ bien déterminé (Klir and Yuan, 1995). Pour expliquer le formalisme prenant un exemple, prenons deux variables d'entrée : taille et poids qui décrivent les règles de surpoids pour une personne. La représentation (définition) de chaque variable suit le quadruple :

$$(\text{Taille}, U = 1, 2, T(X) = \{petit, moyen, grand\}, \mu = \{\mu_{petit}, \mu_{moyen}, \mu_{grand}\})$$

$$(\text{Poids}, U = 45, 200, T(X) = \{maigre, moyen, obese\}, \mu = \{\mu_{maigre}, \mu_{moyen}, \mu_{obese}\})$$

Les termes linguistiques de variable linguistique taille sont:

$$\mu_{petit} = trapzoide(x, \alpha, 1, 1.6, 1.7)$$

$$\mu_{normal} = trapzoide(x, 1.65, 1.75, 1.85)$$

$$\mu_{grand} = trapzoide(x, 1.75, 1.85, 2.00, \beta)$$

Les termes linguistiques de variable linguistique Poids sont:

$$\mu_{maigre} = trapzoide(x, \alpha, 45, 550, 65)$$

$$\mu_{normal} = trapzoide(x, 55, 70, 85)$$

$$\mu_{obese} = trapzoide(x, 70, 85, 200, \beta)$$

Les différents fonctions doivent être définit et valider avec l'aide des experts du domaine étudié afin de proposer un système adéquat.

Comme les variables d'entrées du SIF les variables de sorties doivent être fuzzifié aussi. Pour cela il faut pareillement savoir le nombre de variables de sortie et définir correctement l'univers du discours.

Continuant dans le même exemple, afin d'identifier si une personne souffre d'un surpoids ou

non, on va définir la variable linguistique surpoids. Ces valeurs sont issues de l'indice de masse corporelle(IMC)[1].

$(\text{IMC}, U = 45, 200, T(X) = \{Maigreur, \text{Corpulence normale}, Surpoids\}, \mu = \{\mu_{Maigreur}, \mu_{\text{Corpulence normale}}, \mu$

Qui représente l'univers [10,35], c'est-à-dire le rapport du poids sur la taille $\times$ taille et ses fonctions d'appartenances sont définies comme suit :

$$\mu_{Maigreur} = trapzoide(x, \alpha, 10, 16.5, 19)$$

$$\mu_{\text{Corpulence normale}} = trapzoide(x, 18, 23, 25)$$

$$\mu_{Surpoids} = trapzoide(x, 23, 25, 35, \beta)$$

5.2.3.2 Moteur d'inférence et règles floues

Les règles floues permettent de déduire des connaissances concernant selon les entrées fournies par l'étape de fuzzification. Ces connaissances sont également des termes linguistiques liées au variable linguistique de sortie. Couramment, les règles floues sont collectées des expériences acquises par les experts du domaine. Ces règles simples pouvant être utilisées dans un processus d'inférence floue.

Par exemple, si un expert exprime la règle «si la personne est maigre alors cette personne a un IMC maigreur», alors toute règle floue est généralement de la forme « Si ..... Alors .... » et permet de montrer une relation entre les variables d'entrées et de sortie (Klir and Yuan, 1995). Pour être précis une règle floue r est de la forme suivante :

$$\text{Si } x \text{ est } \gamma \text{ alors } y \text{ est } \delta$$

Où γ et δ sont des termes linguistiques définies dans un univers du discours des variables linguistiques x et y. La première partie de la règle x est γ est la partie prémisse et la deuxième partie de la règle y est δ est la partie conclusion.

Les règles floues, peuvent être simples avec prémisse et conclusion simples ou bien composées, avec la combinaison de plusieurs prémisses de la forme conjonctive suivante :

$$\text{Si } x_1 \text{ est } \gamma_1 \text{ et } x_2 \text{ est } \gamma_2 \text{ et } x_3 \text{ est } \gamma_3 \text{ et } \ldots \text{ et } x_n \text{ n'est pas } \gamma_n \text{ alors } y \text{ est } \delta$$

Considérons à titre illustratif une règle floue : « Si la taille de la personne est petite et le poids est maigre alors l'IMC est maigreur ».

[1] http://www.who.int/fr/

Aujourd'hui, il est cependant possible de constituer une base de règles floues grâce à des méthodes d'apprentissage, sans avoir nécessairement besoin d'un expert humain. Cette stratégie sera décrite plus en détail dans la partie proposition de ce chapitre.

Une fois la base de règles est alimentés par l'ensemble de règles floues, le moteur d'inférence floue détermine les sorties du système à partir des entrées floues issues de la fuzzification des entrées réelles (Klir and Yuan, 1995). Étant donné une base de règles floues, le processus d'inférence consiste à dériver un ensemble flou de sorties à partir de l'agrégation des conclusions de l'ensemble des règles floues (Lin and Lee, 1991).

Dans le cas d'inférence d'une seule règle le degré d'appartenance de la variable linguistique de sortie (ω) est défini comme suit :

$$\mu_{\omega}(x) = \text{ poids } r_1 = min(\mu_{\gamma_1}(x_1), \mu_{\gamma_2}(x_2))$$

Pour expliquer le cas d'une inférence avec une seule règle floue, nous supposons que nous voulons connaître l'IMC d'une personne de poids maigre et qui a une taille petite. Pour cela nous avons deux caractéristiques associées : taille et poids. Ces caractéristiques sont alors nos variables d'entrée et la variable de sortie est IMC. Si nous considérons une personne de taille 1.60 m et qui pèse 50 kg, la règle activée est la suivante :

Si la taille de la personne est petite et le poids est maigre alors l'IMC est maigreur (5.6)

Le degré d'appartenance (μ) pour la variable taille $(x_1 = 1.65)$ et poids $(x_2 = 56)$ en se basant sur les fonctions d'appartenance des variables taille et poids et en appliquant la règle 5.6 est comme suit :

$$x_1 = 1.65 \Rightarrow \mu_{petit} = 0, \mu_{moyen} = 0, \mu_{grand} = 0$$

$$x_2 = 56 \Rightarrow \mu_{maigre} = 0, \mu_{moyen} = 0, \mu_{obese} = 0$$

Le résultat de l'inférence se fait seulement en prenant la valeur la plus petit des prémisses (Voir Figure 5.5) c'est-à-dire que nous obtenons :

$$\text{IMC}_{\mu_{maigreur}} = [min\left(\mu_{petit}(1.65), \mu_{maigre}(56)\right)] = min(0.5, 0.9) = 0.5$$

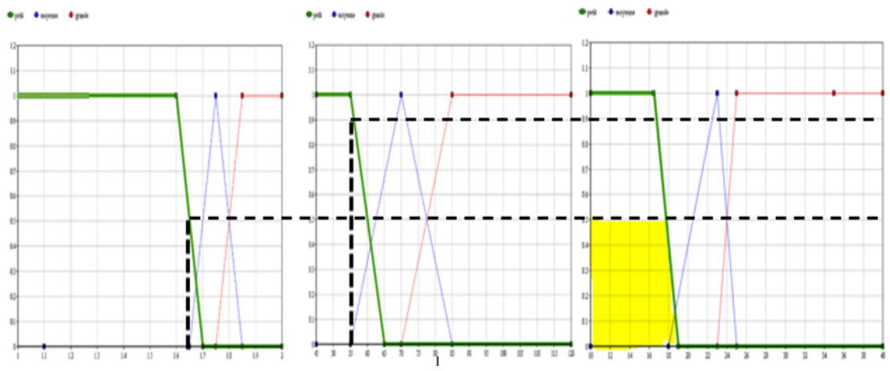


Figure 5.5: Méthode d'inférence d'une règle floue.

Dans le cas de l'inférence de plusieurs règles floues, le moteur d'inférence repose sur les différentes valeurs d'appartenance (μ) liés aux variables linguistiques d'entrée. La définition pour l'activation de plusieurs règles est comme suit :

$$\text{Si } x_1 \text{ est } \gamma_{11} \text{ et } x_2 \text{ est } \gamma_{12} \text{ alors } y \text{ est } \delta_1$$

$$\text{Si } x_1 \text{ est } \gamma_{21} \text{ et } x_2 \text{ est } \gamma_{22} \text{ alors } y \text{ est } \delta_2$$

Dans le cas où δ_1 et δ_2 sont la même valeur de la variable de sortie y, on combine les inférences des 2 règles à l'aide de l'opérateur max. En outre, si δ_1 et δ_2 sont 2 valeurs différentes, chaque règle donne un sous-ensemble flou sur la valeur de sortie y et on agrège les conclusions des 2 règles (Figure 5.8). Cela peut se représenter géométriquement comme on va le voir sur le même exemple :

Nous continuons notre exemple du surpoids avec les valeurs suivantes: la taille 1.8 m et le poids est de 75 kg. Considérons les règles floues suivantes:

- « Si la taille de la personne est moyenne et le poids est moyen alors l'IMC est Corpulence normale » 5.6.

- « Si la taille de la personne est moyenne et le poids est obèse alors l'IMC est Surpoids » 5.7

$$x_1 = 1.8 \Rightarrow \mu_{petit} = 0, \mu_{moyen} = 0.5, \mu_{grand} = 0.5$$

$$x_2 = 75 \Rightarrow \mu_{maigre} = 0, \mu_{moyen} = 0.66, \mu_{obese} = 0.33$$

R_1: « Si la taille de la personne est moyenne et le poids est moyen alors l'IMC est Corpulence normale »

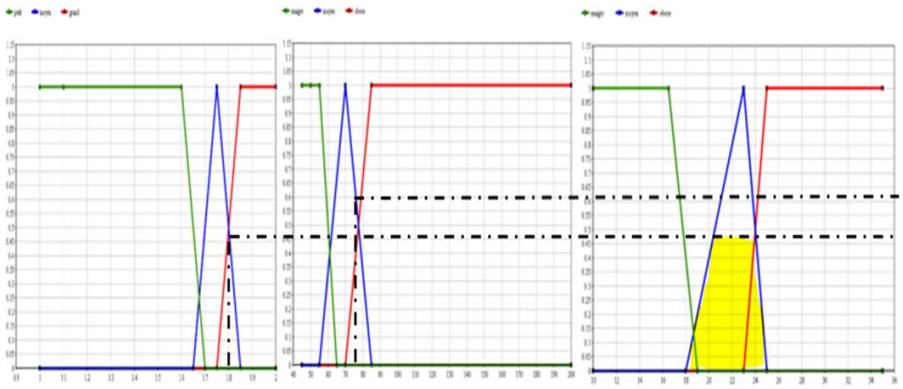


Figure 5.6: Inférence de la première règle floue.

R_1: $\text{IMC}_{\mu_{\text{Corpulence normale}}} = [min\left(\mu_{moyen}(1.8), \mu_{moyen}(75)\right)] = min(0.5, 0.66) = 0.5$

R_2: « Si la taille de la personne est moyenne et le poids est obese alors l'IMC est Surpoids »

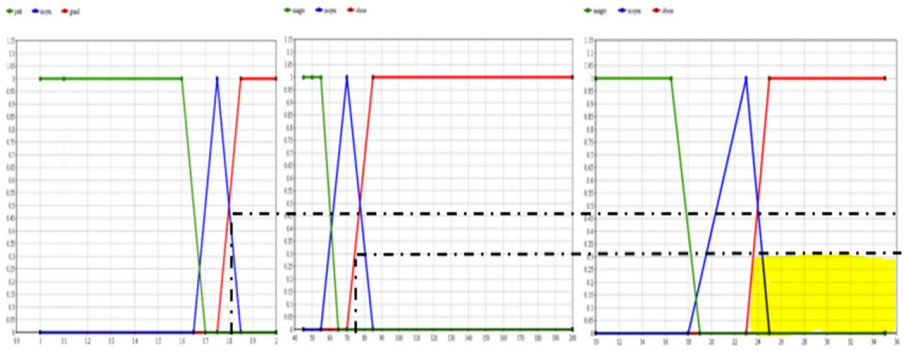


Figure 5.7: Inférence de la deuxième règle floue.

R_2: $\text{IMC}_{\mu_{\text{Surpoids}}} = [min\left(\mu_{moyen}(1.8), \mu_{obese}(75)\right)] = min(0.5, 0.33) = 0.33$

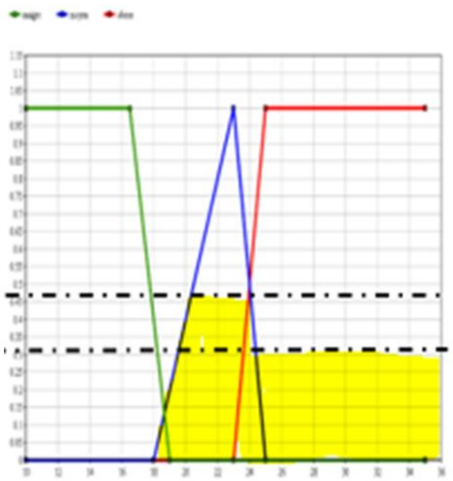


Figure 5.8: Agrégation des deux règles.

Le sous-ensemble obtenu indique que la personne est à la fois d'IMC normal et surpoids. Pour connaître l'IMC exacte de la personne il faut défuzzifier ce résultat pour obtenir y' (Voir Figure 5.8) comme nous l'expliquons dans la section suivante.

5.2.3.3 Déffuzification

La défuzzification permet d'avoir un résultat numérique non flou à partir de la sortie de l'inférence. Cette sortie est un sous-ensemble représentant l'union des conclusions $(\mu_I MC')$. La méthode de défuzzification la plus utilisée pour faire cette transformation est celle de la détermination du centre de gravité. Il existe deux méthodes pour calculer :

- On prend l'union des sous-ensembles flous de sortie et on en tire le centroïde global (calculs très lourds).
- On prend chaque sous-ensemble séparément et on calcul son centroïde, puis on réalise la moyenne de tous les centroïdes.

$$s_A^* = \frac{\int s\mu_A(s)\,\mathrm{d}x}{\int \mu_A(s)\,\mathrm{d}x} \tag{5.7}$$

Où le résultat de la défuzzification est une valeur qui va se transmettre à l'extérieur du système comme résultat de ce mécanisme flou. Les bornes de l'intégrale correspondent alors à la valeur minimale et à la valeur maximale du sous-ensemble IMC, dans notre exemple a = 18 et b = 35. Nous voulons remarquer que l'intégrale du dénominateur donne la surface à défuzzifier, tandis que l'intégrale du numérateur correspond au moment de la surface. La sortie obtenue par la méthode de défuzzification est de 24.33 c'est-à-dire que cette personne a un IMC de 24.33.

Il existe d'autres méthodes pour faire la défuzzification (Lin and Lee, 1991) : méthodes des maximums, somme–prod, moyenne–pondérée, moyenne des maximums, entre autres. Dans notre

thèse on s'intéresse au système d'inférence Mamdani. Eb outre, La méthode du centre de gravité est normalement utilisée avec le mécanisme d'inférence max–min de Mamdani, car cette méthode fournit une interpolation proportionnelle à la taille des sous-ensembles individuels des conséquents.

5.3 CDSS flou pour l'analyse des traces des patients de la maladie coronarienne

Les systèmes d'aide à la décision (SAD) basé logique floue, sont une forme de SAD à base de connaissances (voir chapitre 2). Ils sont à la base de nombreux systèmes complexes dans différents domaines grâce à leur opérationnalité au plus haut degré en matière de productivité et exactitude. Comme on a vu, la logique floue facilite le traitement de l'imprécision dans les SAD. L'approche de la logique floue peut être une approche très précieuse pour décrire le raisonnement humain dans un langage mathématique précis, ce qui donne une force a ce modèle pour représenter les traces médicales (Warren et al., 2000). Dans cette section on va introduire notre contribution pour concevoir un CDSS effective pour l'analyse des traces des patients de la maladie coronarienne.

5.3.1 Descriptions des traces du CAD

CAD comprend une multitude de maladies liées au cœur et le système circulatoire. Troubles cardiovasculaires sont les maladies de l'artère la plus commune coronarienne, qui concernent les artères du cœur, et comprennent, entre autres, l'angine de poitrine, insuffisance cardiaque, infarctus du myocarde (crise cardiaque) et d'AVC du cerveau (AVC) qui se produisent lorsque le cerveau reçoit insuffisance sanguine. L'ensemble des traces est prise à partir du référentiel de l'extraction de données à partir de l'Université de Californie, Irvine (UCI) (Asuncion and Newman, 2007). Robert Detrano MD PhD a recueilli ces données au VA Medical Center. Toutes les expériences publiés liés à l'utilisation d'un sous-ensemble de 14 des 76 attributs de jeu de données Cleveland. L'existence d'un patient de la maladie cardiaque est indiqué dans le « field goal » signifie un nombre entier qui peut prendre n'importe quelle valeur de 0 (pas de présence) à 4. Distribution de classe CAD présent dans 54% et non 46% (voir Table 5.1).

No.	Data set	Attributes	Instances	No. of attributes
1	Cleavland	Age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, Diagnosis	303	14 (6 real, 1 ordered, 3 binary, 3 nominal, 1 binary class)
2	Hungarian		261	
3	Statlog		270	

Tableau 5.1: CAD data sets description

5.3.2 Architecture du Système

Un schéma du système de classification proposé est présenté à la Figure ??. Il est composé de trois niveaux logiques; le niveau de prétraitement comprend des traces de prétraitement à l'aide de la méthode de discrétisation d'intervalles égaux. Cette méthode consistant à changer chaque variable discrète dans un ensemble d'intervalles égaux, ce qui peut nous permettre d'utiliser la méthode SIPINA qui traite uniquement les descripteurs de traces nominaux pour la production graphique de l'induction. Ensuite, l'arbre de décision est générée et les règles sont extraites et alimente la base de règles.

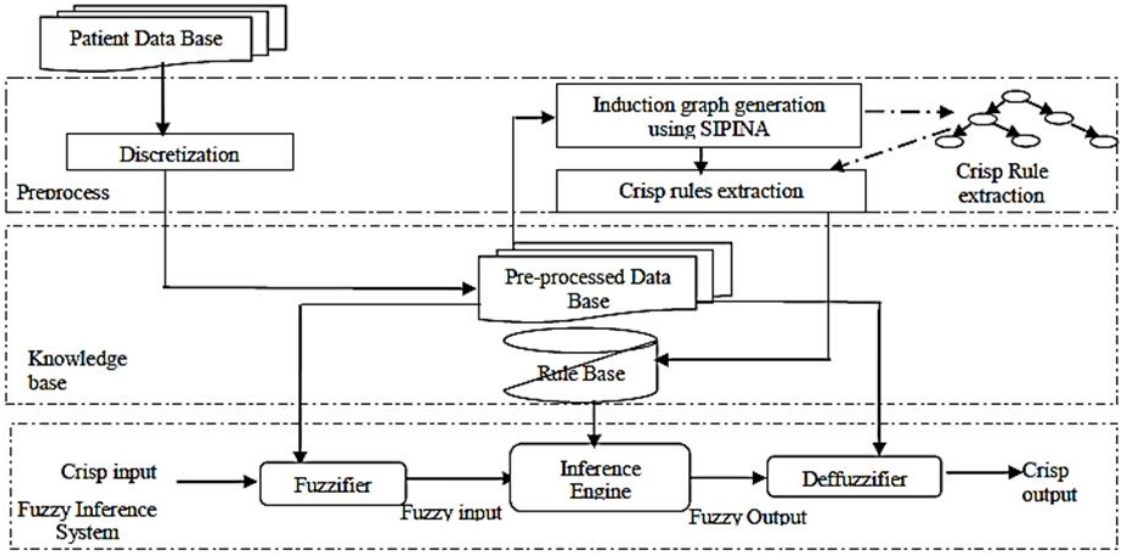


Figure 5.9: Architecture du CDSS flou pour l'analyse des traces du CAD.

En outre, à la fin du deuxième niveau, les règles générées par la méthodologie proposée sont disponibles pour un SAD peut être construit à partir de cet ensemble de règle avec une grande efficacité. L'objectif de la dernière étape est de donner au système la capacité de traiter les aspects de l'incertitude et l'imprécision du domaine médical ce qui résulte un système à base de modèle explicatif. La modélisation floue du système utilise le système d'inférence floue pour Mamdani pour l'analyse des traces des patients du CAD et sa se repose sur trois aspects fondamentaux: le processus de fuzzification, système d'inférence floue (SIF) et le processus de défuzzification.

5.3.2.1 Induction par SIPINA

L'algorithme d'arbre de décision est l'un des algorithmes de fouille de données les plus utilisés. Il s'agit d'un algorithme d'apprentissage par induction par apprentissage supervisé, qui présente l'avantage de la simplicité, de la transparence et l'aptitude à extraire des règles de décision. Il existe plusieurs variantes de méthodes d'arbre de décision dans la littérature comme ID3, la

méthode de C4.5. Dans notre étude, nous faisons usage de la méthode SIPINA, La méthode SIPINA est une heuristique pour la construction d'un graphe d'induction non-arborescent (Zighed et al., 1992). Son principe consiste à effectuer une succession d'étapes de fusion et / ou de fractionnement de nœuds dans le graphe. La particularité de ce processus est de génère l'arbre de décision en introduisant une opération supplémentaire, la fusion, ce qui permet d'obtenir un graphe d'induction. L'opération de fusion existe dans d'autres méthodes (CART, CHAID). Dans tels méthodes, l'opération de fusion ne concerne que les nœuds ayant le même parent, tandis que, SIPINA peut fusionner les nœuds de la structure. Avec Data set du CAD, on commence le traitement symbolique pour la construction du graphe d'induction (méthode SIPINA):

1. Set the measure of uncertainty
2. Set parameters: λ, μ and the initial partition S_0
3. Apply the SIPINA algorithm for going from partition S_i to S_{i+1} and generate induction graph
4. Generate prediction rules

Figure 5.10: Algorithme SIPINA Pour l'extraction des règles.

Chaque nœud S_i fractionne l'espace de traces en deux ou plusieurs sous-espaces selon les valeurs d'attribut. Chaque nœud terminal (feuille) est affecté à une classe d'étiquette et tout chemin de S_0 au nœud terminal dans l'arbre de décision caractérise une règle.

L'objectif de la méthode SIPINA est d'améliorer un critère T_λ, appelés variation de l'incertitude, lors du passage de S_t à S_{t+1}, définis par $\Delta_{t+1} = T_{\lambda_{S_t}} - T_{\lambda_{S_{t+1}}}$.

Dans l'approche proposée, pour le calcul de (S_t) nous utilisons la fonction d'entropie quadratique construite à partir des mesures de l'incertitude:

$$T_\lambda^{(S_t)} = \sum_{j=1}^{K} \frac{n_j}{n} \left(- \sum_{t=1}^{m} \frac{n_{ij}+\lambda}{n_{.j}+m\lambda} \left(1 - \frac{n_{ij}+\lambda}{n_{.j}+m\lambda} \right) \right)$$

Où n_{ij} est la taille de la population venant de la classe c_i, qui est au nœud s_j, n_i. Taille totale de la classe c_i de, $n_{j.}$: Taille totale de nœud s_j, m: Nombre de classes et K: Nombre de nœuds s_j.

La valeur de λ peut être déterminée en utilisant la solution proposée par (Zighed and Rakotomala, 1996) qui définissent un nœud. La valeur de λ peut être déterminée pour toutes les distributions possibles T_k de μ individus sur les classes m :

$$\lambda = \max \left(\lambda(m-1) \frac{2\mu^2 + 2\mu + m\lambda + 2\mu m\lambda}{(\mu + m\lambda)^2(\mu + 1 + m\lambda^2)} \right)$$

Avec

$$\mu = n_s = -\log(\alpha_0)\left(\frac{n}{n_c}\right)$$

Où $n = ns + nc$, ns est le nombre de traces appartenant a la classe de la classe, et nc est le nombre traces qui n'appartient pas à la classe. $\alpha_0 = 0.05$. La table 5.2 montre un échantillon de règles extraites utilisant la méthode SIPINA.

IF (sex is Male and trestbps and >105 fbs = 1 and thalach $\leqslant$ 173 and oldpeak $>$ 0.6 and ca $=$ 2)	THEN	CAD
IF (cp >1 and chol $\leqslant$ 318 and exang = 1 and oldpeak $<$ 1.6 and slope $\leqslant$ 2 and ca = 0)	THEN	NORMAL
IF (fbs = 1 and restecg $\leqslant$ 1 and exang = 1 and oldpeak >1.6 and thal $\leqslant$ 2)	THEN	CAD
IF (cp >1 and oldpeak $>$ 1.8 and slope $\leqslant$ 2 and ca $\leqslant$ 0) and thal $>$ 2)	THEN	CAD
IF (Age >45 and cp >1 and exang = 1 and oldpeak $\leqslant$ 1.8 and ca = 0)	THEN	NORMAL
IF (trestbps >110 and chol $\leqslant$ 269 and exang = 1 and oldpeak $\leqslant$ 2.40)	THEN	NORMAL
IF (restecg $\leqslant$ 1 and exang = 1 and oldpeak $\leqslant$ 1.6 and thal $\leqslant$ 2)	THEN	NORMAL
IF (Age >55 and sex is Female and chol $>$ 244 and thal $\leqslant$ 2)	THEN	CAD
IF (trestbps >108 and ca $\leqslant$ 2 and thal $>$ 2)	THEN	CAD
IF (Age $\leqslant$ 55 and thal $\leqslant$ 2)	THEN	NORMAL

Tableau 5.2: Échantillon de règles extraites utilisant SIPINA, Thal = Thallium Scan, CP = Chest Pain type, trestbps = Blood Pressure, OldPeak = Old Peak, fbs = Fasting blood sugar, thalach = Heart Rate, slope = the slope of the peak exercise ST segment,ca = number of major vessels by flourosopy, exang = exercise induced angina, restecg = Resting ECG, Chol = Serum Cholesterol

5.3.2.2 Système d'inférence Flou

Après l'induction de l'arbre de décision usant la méthode SIPINA, nous construisons un classificateur à base de règles. Les règles du modèle sont classiques. La prochaine étape est la modélisation floue qui transforme le modèle des règles classique en modèle des règles floues en utilisant une fonction d'appartenance floue Afin de de développer un CDSS en utilisant la logique floue pour évaluer le niveau de risque en analysant les traces du CAD.

La logique floue introduit par (Zadeh, 1965) est la redécouverte de la logique à valeurs multiples. Une logique floue, règles floues, moteur d'inférence floue et defuzzifier existent dans un modèle de logique floue. Logique floue: Tout d'abord, les variables linguistiques floues, termes linguistiques floues et fonctions d'appartenance sont utilisés pour transformer un ensemble classique recueilli des traces d'entrée dans un ensemble flou. Cette phase est reconnue comme fuzzification. Base de règles floues: Les règles floues qui sont importantes pour tout système flou sont bien définis après les entrées sont fuzzified. Moteur d'inférence: Ensuite, un ensemble de règles définies dans

la base de règles floues est utilisé comme base pour l'interprétation et à l'aide de sorties floues sont générés. Defuzzifier: Un ensemble flou est utilisé comme entrée pour les fonctions du processus de défuzzification et la fonction d'appartenance de la variable de sortie est utilisée pour produire la sortie en valeur réel (voir section 2).

Processus de fuzzification Les paramètres d'entrée dans notre système expert flou ont été fuzzifié. Les facteurs de risque décrits dans le chapitre 1 sont les paramètres d'entrée pour notre système actuel : l'âge, Thallium scan (thal), chest pain(cp), serum cholesterol (chol), ST depression induced by exercise relative to rest(oldpeak), maximum heart rate achieved (thalach), the slope of the peak exercise ST segment (slope) et resting blood pressure (trestbps). La représentation linguistique des paramètres d'entrée est indiquée dans la table 5.3. Les valeurs de ces variables floues sont définies par degré d'appartenance, qui est déterminé par la fonction d'appartenance. Parmi les différentes fonctions d'appartenance, nous considérons ici fonction d'appartenance triangulaire où deux points sont les mêmes sommets. L'équation de la fonction d'appartenance triangulaire est définie comme suit:

$$Triangular : f(x; a, b, c) = \begin{cases} 0, & x \leqslant a \\ \frac{x-a}{b-a}, & a \leqslant x \leqslant b \\ \frac{c-x}{c-b}, & b \leqslant x \leqslant c \\ 0, & c \leqslant x \end{cases}$$

Où les paramètres a et b représentent les limites inférieure et supérieure respectivement, tandis

Linguistic Variables	Range	Fuzzy Set
Age	0-45	Young
	40-60	Mature
	55-77	Old
Trestbps	0-118	Hypotension
	100-182	Normal
	145-185	Hypertension Crisis
	155-250	Hypertension
Serum Cholesterol	0-200	Normal
	170-270	Low
	250-685	High
Thalach	0-141	bradycardia
	115-185	Normal
	150-400	Tachycardia
OldPeak	0-4.25	low
	2.30-6.20	High

Tableau 5.3: Sous ensemble flou des différents descripteurs du CAD

que c localise le sommet du triangle. Comme mentionné ci-dessus, la section importante du

modèle flou est les fonctions d'appartenance floues de chaque descripteur. Pour ce modèle flou, il existe 13 descripteurs d'entrée et un de sortie (voir table 5.3 et Figure 5.11).

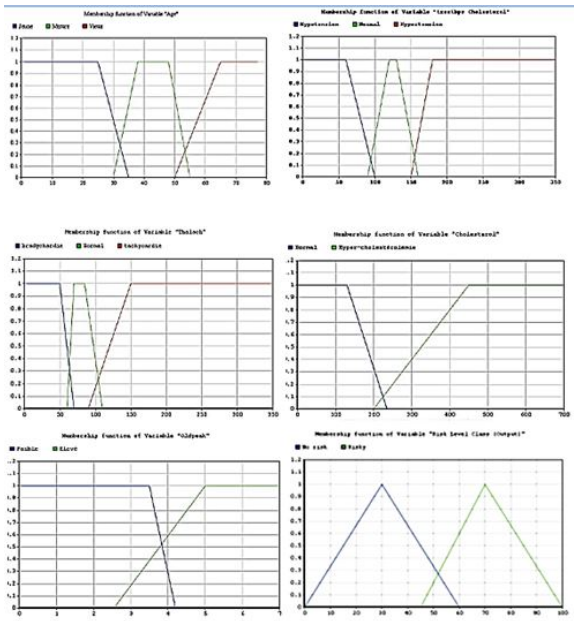


Figure 5.11: Fonctions d'appartenances relatives aux descripteurs du CAD.

Système d'inférence flou et defuzzification La base de connaissances de notre système a été développée avec l'aide des experts du domaine en se basant sur l'expérience antérieure et de l'intuition médicale. Après plusieurs séries tests avec les experts pour la validation de la base de règle floue. Tout conflit entre experts a été résolu par le consensus. Il existe de nombreuses techniques pour la représentation des connaissances, graph conceptuel, ontologies, les règles de production, etc. Le présent module utilise des règles de production floues pour représenter la connaissance. Les règles de production sont écrites dans le format <IF (condition) THEN (conclusion)>. Dans le système flou la partie condition de la règle et la conclusion sont des variables floues. La méthode des centres de gravité est utilisée pour la defuzzification. La sortie de notre CDSS est le risque d'atteindre un CAD relative à une trace d'un patient.

5.4 Évaluation du CDSS flou pour les patients du CAD

Après la régulation des fonctions d'appartenances et la production de la base de règles floues, les règles ont été obtenues à partir d'un ensemble de traces de CAD de 303 cas pour lesquels 164 cas étaient en bonne santé et 139 cas étaient condition de maladie cardiaque. Pour tester le modèle flou intégré, une validation 10-crosée a été utilisée. Dans notre thèse, les traces originales sont divisées en 10 sous-ensembles de traces et pour chaque étape de validation, un seul sous-ensemble est utilisé comme traces de test et les sous-ensembles restants sont utilisés comme traces

Result of the diagnostic test		Physician diagnosis	
		Negative	Positive
Classifier Result	Negative	159	5
	Positive	13	126

Tableau 5.4: Matrice de confusion du CAD

d'apprentissage. Ce processus est répété jusqu'à ce que tous les sous-ensembles de traces soient utilisés comme traces de test. Le CDSS flou créé a été évaluée et sa performance a été donnée comme matrice de confusion. Le nombre de cas classifiés correctement est présenté dans les éléments diagonaux de la matrice de confusion. Dans la première rangée, le premier élément indique le nombre de cas appartenant à un cas sain et classés par la FIS en bonne condition. Le deuxième élément dans la deuxième rangée illustre les traces appartenant à une maladie cardiaque et classés par la FIS qu'elles présentent une maladie cardiaque (Table 5.4).

Metric	*Proposed System results*
Sensitivity	*92.44*
Specificity	*96.18*
Accuracy	*94.05*

Figure 5.12: performance du CDSS proposé en matière de prévision des risques du CAD.

Les paramètres d'évaluation sont calculés les ensemble de traces de tests et le résultat obtenu est illustré dans la figure 5.12.Par conséquent, l'objectif de l'étude comparative est de trouver l'efficacité de la proposition de CDSS face aux différents systèmes existants dans la littérature. En effet tableau 5 révèle que la méthode proposée permet d'obtenir la plus grande exactitude en matière de prédiction des patients du CAD.

En conséquent, l'objectif de l'étude comparative est de trouver l'efficacité de la proposition de CDSS face aux différents systèmes existants dans la littérature. En effet la table 5.5 révèle que la méthode proposée permet d'obtenir la plus grande exactitude en matière de prédiction des patients du CAD. Ensuite, l'ensemble de traces de tests est donnée au système proposé, et aux différents systèmes de la littérature pour prédire le niveau des risques des maladies cardiaque. Ici, les résultats de la prédiction sont analysés par la méthode de 10- fois la validation croisée et les valeurs de sensibilité, de spécificité et d'exactitude correspondantes sont calculées. Les valeurs sont ensuite illustrées en Figure 5.13 . Lors de l'analyse graph ci, le système proposé surperformé deux autres méthodes. Dans l'exactitude globale, notre proposition (CDSS Flou) atteint 94,05%, tandis que l'exactitude du système expert flou basé PSO et le mécanisme de résolution Fuzzy sont 93,27% et 91,48% respectivement ce qui prouve l'efficacité de notre système par rapport aux autres contributions du domaine.

Auteur	Méthode	Accuracy (%)
(Chen et al., 2011)	Artificial Neural Network	80.00
(Chauraisa and Pal, 2013)	CART	83.40
(Shouman et al., 2012)	Naive Bayes	83.50
(Marateb and Goudarzi, 2015)	MLR+NFC	84.00
(Sundar et al., 2012)	WAC	84.00
(Bashir et al., 2015)	BagMOOV	84.16
(Kahramanli and Allahverdi, 2008)	Hybrid neural network system	87.00
(Polat et al., 2006)	Hybrid neural network system	89.01
(Das et al., 2009)	Neural network ensemble	89.01
Proposed System	Fuzzy CDSS	94.05

Tableau 5.5: Comparaison des résultats du système proposé avec les recherches similaires.

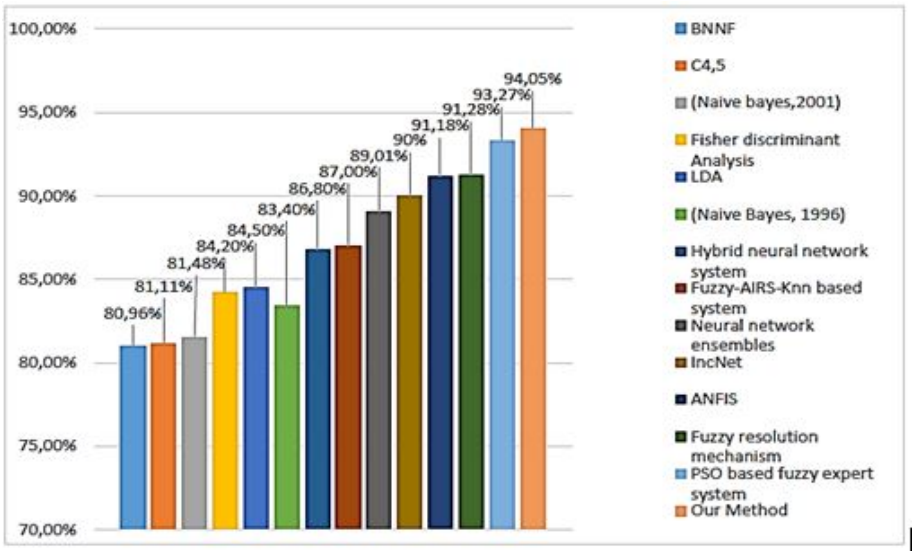


Figure 5.13: Validation du CDSS flou vis-à-vis les différents systèmes existants.

5.5 Conclusion

Dans ce chapitre, un CDSS flou basé sur la méthode SIPINA a été développé afin de prédire les risques des maladies cardiaques relatives aux traces des patients. Avec cette approche proposée, 94,05% classification correcte sur l'ensemble de test a été achevé. La découverte des attributs significatifs et des règles classique a été réalisée en utilisant l'algorithme d'arbre de décision (SIPINA).

L'importance de la découverte automatique de règles floues est une étape importante et pertinente

et la validation de ces règles par les experts est primordiale. Les principaux avantages de notre CDSS comme un outil d'extraction de connaissances sont les suivants: (1) un petit nombre de règles sont obtenus (2) les règles obtenues peuvent être facilement interprétés.

Dans les études futures, le CDSS conçu peut être généralisée à être utilisé pour différentes pathologies médicales. Nous allons essayer d'améliorer constamment pour la prédiction des différents niveaux de risque de CAD. Une des œuvres futures liées à notre CDSS est d'integrer un système de recommandation afin de minimiser le risque.

Conclusion Générale et Perspectives

Le sujet de cette thèse traite deux tâches de l'analyse des traces des patients: l'extraction des descripteurs pertinents pour l'analyse des traces et l'analyse des traces des patients pour l'aide à la décision médicale.

Dans le cas de : l'extraction des descripteurs pertinents pour l'analyse des traces, nous avons proposé et développé un algorithme, nommée GA-NB, capable d'améliorer la performance des différents classificateurs. L'objectif de cette contribution est de sélectionner un minimum de descripteur de traces sans que la performance prédictive ne soit affectée.

La partie la plus intéressante au niveau de la recherche est la proposition de l'algorithme génétique comme un support pour l'exploration des traces pour chercher le meilleur sous-ensemble de descripteurs représentatifs; premièrement, l'algorithme génétique est une méthode d'(optimisation aléatoire qui vise à trouver une solution optimale sans explorer tout l'espace et deuxièmement, il est considéré comme un outil de recherche et d'extraction des traces les plus pertinents pour la classification.

Les performances de classification avec GA-NB sont évaluées avec trente-quatre data set des référentielles les plus connus (UCI/KEEL) pour différents domaine: médicale, reconnaissance de forme, imagerie, biologie, etc. Enfin l'algorithme a été appliqué sur un ensemble de traces des patients du CAD. Les résultats sont comparés avec les différentes méthodes existantes dans la littérature. En utilisant le principe génétique (par chromosomes) pour la représentation des traces et la sélection des descripteurs pertinents pour la classification, nous avons montré dans le cas de la majorité des data set que notre algorithme améliore non seulement l'exactitude de la classification mais aussi réduit le cout en temps et de calculs.

Nous pouvons dire que les algorithmes génétiques représentent un outil très puissant pour la réduction de dimensionnalité des traces. Pour cela il faut d'abord définir les variables d'entrée et de sortie du système, ensuite, le choix du codage et de la méthode de sélection. Le premier choix est déterminé par plusieurs facteurs à savoir le type des variables à et l'exactitude recherchée. Quant au choix de la méthode de sélection, il faut tenir compte du nombre de critères de performance défini, s'il s'agit d'un seul critère, la méthode de sélection par tournoi peut être considérée dans un premier coup qui est le cas dans le problème de réduction de dimensionnalité. Nos principales contributions dans ce travail sont résumées comme suit :

- Nous proposons et développons une technique de classification qui améliore les performances de l'algorithme Naïve Bayes tout en réduisant le temps de classification;
- En se basant sur l'algorithme génétique, nous proposons un système de sélection de descripteurs qui nous permet de réduire la dimensionnalité des traces.
- Nous améliorons les performances de classification NB en sélectionnant uniquement les traces pertinentes ;

Et enfin, nous évaluons cet algorithme pour la prédiction du CAD d'une manière générale. La deuxième contribution de cette thèse correspond à l'introduction d'une approche pour l'analyse de traces des patients du CAD. Cette approche utilise la logique floue pour construire un système d'aide à la décision clinique CDSS. Pour mieux guider ce parcours nous introduisons une méthode de classification pour l'extraction automatique des règles classiques. Les connaissances extraites à partir de la méthode de classification SIPINA permet de construire une base de connaissance classique. La logique floue a été utilisée pour représenter l'imprécision et l'incertitude du raisonnement médicale. Un ensemble de règles classiques a été transformé et valider par les experts domaine. Pour cela, nous avons développé un CDSS pour l'analyse des traces du CAD capable d'analyser les propriétés des traces des patients du CAD. Ce Système intègre la logique floue qui nous permet de :

- Une modélisation flou des traces pour représenter les règles floue et
- Un moteur d'inférence floue pour la prédiction du risque du CAD.

Les résultats d'évaluation du système sur une base de traces du CAD sont très satisfaisants. Nous avons obtenu un taux d'exactitude, taux sensitivité et taux de specifité égaux à 94.04%, 92.44% et 96.18 % respectivement. Cependant, ce travail nécessite toujours des améliorations selon plusieurs directions :

- Acquérir d'autres base des traces des patients Algérien du CAD pour évaluer la qualité d'extraction d'une manière plus concise;
- Améliorer la méthode de sélection de traces en lui ajoutant l'aspect parallélisme;
- Cette étude peut être aussi poursuivie en considérant la tâche de recommandation de thérapies ce qui nécessiterait d'intégrer dans notre système une base de recommandations en l'occurrence des concepts médicaux et la relié éventuellement avec les classes prédéfinies.
- Faire participer des spécialistes du domaine médical dans la construction de cette base de recommandations pour éviter toute erreur de marquage.
- Et comme perspectives à long terme, il serait intéressant d'intégrer notre CDSS dans une application réelle comme la fouille de traces médicales.

Annexe A

Annexe A

A.1 Base UCI

Nous avons en premier lieu réalisé des expérimentations sur 10 bases de traces UCI, que nous allons décrire :

A.1.1 Arrhythmia

Cette base de traces contient 279 attributs, cette base de trace est décrite par un ensemble de 206 attributs de type continu et 73 attributs nominaux. Le but de cette base est de déterminer la présence ou l'absence de l'arythmie cardiaque[1]. La classe 01 présente un ECG normal, les classes entre 02 et 15 présentes différentes classes d'arythmie et la classe 16 regroupent les traces des patients non classées. Les noms et les identifiants des traces de ces patients ont été supprimés utilisant un algorithme d'anonymisation (Güvenir et al., 1997).

A.1.2 Hypothyroïdie

La base de trace Hypothyroïdie contient les résultats du suivi et le traitement des patients suspects d'avoir des troubles thyroïdiennes[2]. La distribution des patients est 92.5% sont negatifs, 5% présente l'hypothyroïdie au premier stade, et 2.5% ont un type d'hypothyroïdie compensée. Cette base de traces présente des données hétérogènes avec un grand nombre de descripteurs (Quinlan et al., 1987b).

[1] https://archive.ics.uci.edu/ml/datasets/Arrhythmia
[2] https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease

A.1.3 Letter

L'objective est d'identifier chaque grand nombre d'écrans rectangulaires de pixels noir et blanc comme l'une des 26 lettres majuscules de l'alphabet [3]. Les images des différentes lettres sont fondées sur 20 polices différentes et chaque lettre dans ces 20 polices était déformée de façon aléatoire pour produire un fichier de 20 000 traces uniques. Chaque trace a été converti en 16 attributs numériques primitifs (moments statistiques et les chiffres de bord) qui ont ensuite été mis à l'échelle pour s'adapter à une gamme de valeurs entières de 0 à 15 (Frey and Slate, 1991).

A.1.4 Lymphographie

Cette base de données Lymphographie a été obtenue du Centre universitaire de médecine, Institut d'oncologie, Ljubljana, Yougoslavie. Il y a 148 traces des patients au total sans valeurs manquantes [4]. Il y a 18 attributs à valeur numériques qui sont énumérés comme suit (machine-apprentissage des bases de données, 2007):

1. lymphatiques: normal, arqué, déformé, déplacé;
2. bloc de affere: non, oui;
3. Bloc de lymphe c — valeur 1 pour non & 2 pour oui;
4. Bloc de lymphe s — valeur 1 pour non & 2 pour oui;
5. By pass — valeur 1 pour non & 2 pour oui;
6. extravasâtes — expulsé à partir d'un récipient: non, oui;
7. Régénération —: non, oui;
8. Absorption précoce: non, oui;
9. Les ganglions lymphatiques dimension - [0-3];
10. Les ganglions lymphatiques élargissement - [1-4];
11. Type de changements dans la lymphe: haricot, ovale, ronde;
12. défectuosité du nœud: non, lacunaire, lacunaire marginale, lacunaire central;
13. changements du nœud: non, lacunaire, lacunaire marginale, lacunaire central;
14. Changements dans la structure - la structure du système lymphatique;
15. Formes spéciales - aucun, calices, vésicules;
16. dislocation du nœud - : non, oui;
17. Exclusion du nœud - : non, oui;
18. Nombre de nœuds - varie de 0 à 80.
19. Il existe quatre valeurs pour la classe diagnostic: normal, métastases lymphatiques, maligne et fibrose.

[3] https://archive.ics.uci.edu/ml/datasets/Letter+Recognition
[4] https://archive.ics.uci.edu/ml/datasets/Lymphography

A.1.5 Multifeat

Multifeat, est une base de traces qui se compose de chiffres manuscrits extraites d'une collection de cartes de services publics néerlandais. 200 modèles par classe (pour un total de 2000 modèles) ont été numérisés en images binaires[5]. Ces chiffres sont représentés en termes de six traces suivantes (Van Breukelen et al., 1998):

1. mfeat-fou: 76 coefficients de Fourier des formes de caractères;
2. mfeat-fac: 216 corrélations profil;
3. mfeat-kar: 64 coefficients Karhunen-Love;
4. mfeat-pix: 240 moyennes de pixels en 2 x 3 fenêtres;
5. mfeat-zer: 47 moments de Zernike;
6. mfeat-mor: 6 variables morphologiques

A.1.6 Mushroom

Cet ensemble de traces comprend des descriptions d'échantillons hypothétiques correspondant à 23 espèces de champignons à lamelles dans la famille Agaricus et Lepiota (pp. 500-525) [6]. Chaque espèce est identifiée comme définitivement consommable, certainement toxique, ou de comestibilité inconnue et non recommandée. Cette dernière classe a été combinée avec celui toxique. Le Guide indique clairement qu'il n'y a pas de règle simple pour déterminer la comestibilité d'un champignon (Schlimmer, 1987).

A.1.7 Sick

Une base de traces des patients de la thyroïde fournis par l'Institut « Institut New South Wales », Sydney, Australie (1987)[7].

A.1.8 Soybean

SoyBean est une célèbre base de traces de la maladie de Michalski [8]. Il y a 19 classes, seules les 15 premières qui ont été utilisées dans des travaux antérieurs. Les quatre dernières classes sont injustifiées par les données, car ils ont si peu d'exemples. Il y a 35 attributs catégoriques, certains nominales et certains continues. La valeur "dna" signifie «does not apply» (Michalski and Chilausky, 1980).

[5] https://archive.ics.uci.edu/ml/datasets/Multiple+Features
[6] https://archive.ics.uci.edu/ml/datasets/Mushroom
[7] http://tunedit.org/repo/UCI/sick.arff
[8] https://archive.ics.uci.edu/ml/datasets/Soybean+(Large)

A.1.9 Splice

Les jonctions d'épissage sont des points sur une séquence d'ADN à laquelle l'ADN superflu est retirée pendant le processus de création de protéines dans les organismes supérieurs. Le problème posé dans cette base de traces d'ADN est de reconnaître, étant donné une séquence d'ADN, les limites entre exons (les parties de la séquence d'ADN conservé après épissage) et les introns (parties de la séquence d'ADN qui sont épissé)[9]. Ce problème se compose en deux sous-tâches: reconnaître les frontières exons / introns (appelés sites de l'assurance-emploi), et reconnaissant intron / limites d'exons (sites IE). (Dans la communauté biologique, frontières IE sont renvoyés à un «accepteurs » si les frontières d'assurance-emploi sont appelés «donateurs».

Cette base de traces d'ADN a été développée pour aider à évaluer un algorithme «hybride» d'apprentissage (KBANN) qui utilise des exemples pour affiner la connaissance inductive préexistante avec l'utilisation de la validation croisé sur 1000 traces d'ADN choisis au hasard parmi l'ensemble des 3190 (Noordewier et al., 1991).

A.1.10 Waveform

Waveform, est un problème à trois classes dans une espace dimensionnelles à 40 attributs [10]. Les 21 premières attributs de chaque catégorie représentent une onde générée à partir d'une combinaison de deux des trois formes d'onde décalées triangulaires, plus le bruit gaussien de moyenne 0 et de variance 1. Les derniers 19 attributs sont tous bruit gaussien. Il peut être prouvé que la dimension discriminante intrinsèque de ce problème est 2. La base Waveform contient 5000 traces (Breiman et al., 1984).

[9] https://archive.ics.uci.edu/ml/datasets/Molecular+Biology+%28Splice-junction+Gene+Sequences%29
[10] https://archive.ics.uci.edu/ml/datasets/Waveform+Database+Generator+%28Version+2%29

Annexe B

Annexe B

B.1 Base KEEL

KEEL est un référentiel qui comprend un ensemble de bases de traces des partitions sous la forme KEEL et montre des résultats d'algorithmes de ces bases des traces ensembles. Ce dépôt fournit aux chercheurs des bases de traces pour faire la comparaison de leurs modèles avec ceux déjà existants.

B.1.1 Optical Digits

Cette base de trace appelé reconnaissance optique de chiffres Manuscrit, 2 à 64 descripteurs et 5620 traces. Il consiste à identifier des chiffres de 0 à 9, il est donc un ensemble de traces de classe multi-valeurs[1].

Les créateurs de cette base de traces avaient utilisés les programmes de prétraitement mis à disposition par le NIST (National Institute of Standards and Technology) pour extraire les bitmaps normalisés des chiffres manuscrits à partir d'un formulaire préimprimé effectué par un total de 43 personnes. Les bitmaps 32x32 sont divisées en blocs non chevauchantes de 4x4 et le nombre de pixels sur plus sont dans chaque bloc. Ceci génère une matrice d'entrée de 8x8 où chaque élément est un nombre entier dans l'intervalle 0..16. Cela a réduit la dimensionnalité et donne invariance aux petites distorsions (Xu et al., 1992).

B.1.2 anneal.ORIG

anneal.ORIG est une base de traces de recuit, cette base contient 689 traces décrites par différents types d'attributs (Xu et al., 1992)

[1] http://sci2s.ugr.es/keel/dataset.php?cod=199

B.1.3 Colic

Colic est une base de traces pour l'identification des maladies de chevaux[2]. La tâche consiste à déterminer si la lésion du cheval était chirurgicale ou non. C'est une version modifiée de l'ensemble de données d'origine, où seulement un attribut de classe est considéré (Surgical_lesion) (Greensmith, 2003).

B.1.4 cylinder-bands

Un problème de classification de Rotogravure, où la tâche est de déterminer si une pièce donnée est une bande de cylindre[3] (Evans and Fisher, 1994).

B.1.5 kdd synthetic control

Ces données se compose de cartes de contrôles générés automatiquement. Cette base de données contient 600 exemples de cartes de contrôle de synthèse générés par le procédé dans Alcock et al. (1999). Il y a six classes différentes de cartes de contrôle(voir Figure B.1).

[2] http://sci2s.ugr.es/keel/dataset.php?cod=180#sub1
[3] http://sci2s.ugr.es/keel/dataset.php?cod=89

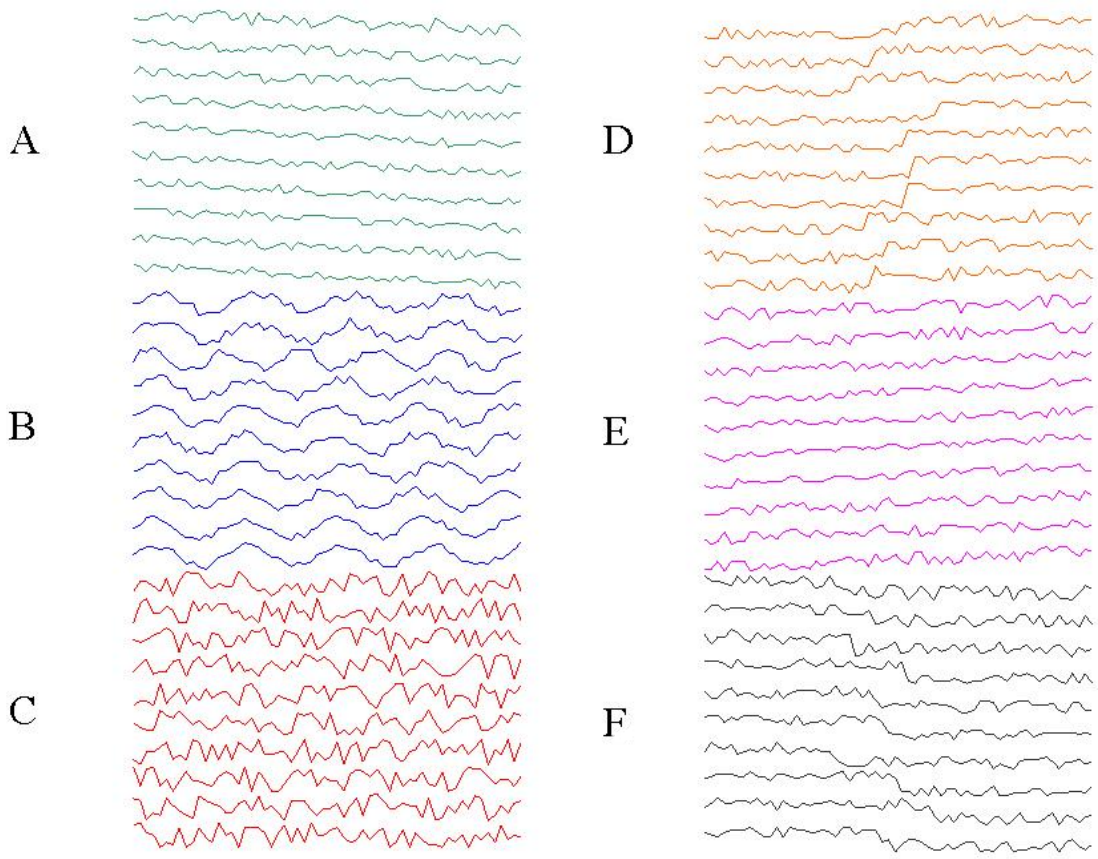


Figure B.1: Exemples de différentes cartes de contrôles, (A) normal,(B) cyclique, (C) tendance à la hausse, (D) tendance à la baisse, (E) déplacement vers le haut, (F) tendance de basse

B.1.6 Autos

Cet ensemble de traces de données relatives au automobiles comprend trois types d'entités: (a) la spécification d'une auto en termes de ses caractéristiques différentes, (b) sa cote d'assurance risque attribué, (c) Les différents accidents pendant l'utilisation. La deuxième note correspond à la mesure dans laquelle l'auto est plus risquée que son prix indiqué. Les voitures sont initialement assignés à un symbole de facteur de risque associé à son prix. Ensuite, si elle présente plus de risque (ou moins), ce symbole est réglée en déplaçant vers le haut (ou le bas) de l'échelle ce processus est appelé "symboling". Une valeur de 3 indique que l'auto est risquée, -3 que l'automobile ne présente quasiment aucun risque. Le troisième facteur est le paiement de la perte moyenne par an pour l'assurance véhicule. Cette valeur est normalisée pour toutes les autos dans une classification de taille particulière (deux portes petite, breaks, sports / spécialité, etc ...), et représente la perte moyenne par voiture par année. En outre, plusieurs attributs dans la base de traces pourraient être utilisés comme attribut de classe (Kibler et al., 1989).

B.1.7 Chess

Un ensemble de traces représentant une fin de match d'échecs. La base contient des exemples de positions d'échecs décrits que par les coordonnées des pièces sur le plateau. Les connaissances de base sous la forme de différences de lignes et de colonnes est également fourni. Les fins de parties d'échecs sont des domaines complexes qui sont dénombrables. Les valeurs de la théorie des jeux stockés représentent si les positions des pièces d'échecs permettent de gagner pour chaque côté, ou également la profondeur le nombre de coups en proposant le meilleur coup optimale. La tâche est de déterminer si Le Blanc peut gagner ou non[4] (Bain and Muggleton, 1994).

B.1.8 Coil2000

cet ensemble de traces utilisé contient des informations sur les clients d'une compagnie d'assurance, ces informations sur les clients se compose de 86 des variables et comprend les données d'utilisation des produits et les données sociodémographiques[5]. Les traces ont été fournies par la société de fouille de données néerlandaise Sentient Machine Research et est basé sur un problème d'entreprise du monde réel. L'ensemble de traces contient plus de 5000 descriptions de clients, y compris les informations (Van Der Putten and Van Someren, 2004).

B.1.9 Connect-4

Cette base de données contient toutes les positions juridiques dans le jeu de Connect-4 pour une grille de 6x7, où aucun joueur n'est encore gagné, et dans lequel le prochain mouvement n'est pas forcé [6] . Ainsi, chaque attribut contient une valeur nominale qui décrit si une position donnée est nulle ou si elle a été occupée par un joueur. La tâche est de prédire quel joueur est susceptible de gagner le match (Burton and Kelly, 2006).

B.1.10 Dermatology

Le diagnostic différentiel des maladies érythémato-squameuses est un réel problème dans la dermatologie. Les patients ont été évalués cliniquement premier avec 12 fonctions. Ensuite, des échantillons de peau ont été prélevés pour l'évaluation des 22 caractéristiques histopathologies [7](Güvenir et al., 1998).

[4] http://sci2s.ugr.es/keel/dataset.php?cod=187
[5] http://sci2s.ugr.es/keel/dataset.php?cod=91
[6] http://sci2s.ugr.es/keel/dataset.php?cod=193
[7] http://sci2s.ugr.es/keel/dataset.php?cod=60

B.1.11 Est-West

Est-West est à l'origine un problème ILP. Le problème est de prédire si un train est vers l'est ou vers l'ouest. Un train contient un nombre variable de voitures qui ont différentes formes et transportent des charges différentes [8].

B.1.12 KDD CUP 99

la tâche est de détecter les intrusions réseau pour protège un réseau contre les utilisateurs non autorisés. L'objectif est de construire un modèle prédictif de détection d'intrusion en se basant sur une base de traces du trafic réseau capable de distinguer entre mauvaise connexion, appelées intrusion ou attaque, et bonne connexion appelée normale. Pour changer les types des bases de traces, on a utilisé deux versions de cette base :

- kddcup-buffer_overflow_vs_back[9] , Une version déséquilibrée de l'ensemble de traces du KDD cup 99, où les exemples positifs appartiennent au buffer_overflow et les exemples négatifs classe Back.
- kddcup-guess_passwd_vs_satan , une autre version desequillibrée de la base de traces KDD cup 99. Dans cette base les exemples positifs sont representés par la classe guess_passwd et les ensembles négatifs à la classe satan.

B.1.13 Movement_libras

L'ensemble de traces contient 15 classes contenant 24 traces chacune, où chaque classe est un type de mouvement de la main dans LIBRAS (officielle langue de signal brésilien)[10]. Dans le prétraitement de la vidéo, une normalisation de temps est effectuée en sélectionnant l'une de 45 trames vidéo, en fonction à une distribution uniforme. Dans chaque image, les barycentres pixels identifient les objets segmentés (la main), ce qui composer la version discrète de la courbe F avec 45 points. Toutes les courbes sont normalisées dans l'espace unitaire. Afin de préparer ces mouvements pour l'analyse, une opération de mise en correspondance dont chaque courbe F est mappée dans une représentation à 90 attributs, pour représenter les coordonnées de mouvement (Dias et al., 2009).

B.1.14 Musk

Cet ensemble de traces décrit un ensemble de 102 molécules dont 39 sont jugés par des experts humains d'être muscs et les 63 molécules restants sont jugés non-muscs. Le but est d'apprendre

[8] http://sci2s.ugr.es/keel/dataset.php?cod=167
[9] http://sci2s.ugr.es/keel/dataset.php?cod=1316
[10] http://sci2s.ugr.es/keel/dataset.php?cod=165

à prévoir si de nouvelles molécules seront muscs ou non-muscs [11]. Toutefois, les 166 attributs qui décrivent ces molécules dépendent de la forme exacte, ou conformation, de la molécule. Une seule molécule peut adopter de nombreuses formes différentes. Pour générer cet ensemble de données, toutes les conformations de faible énergie des molécules ont été générées pour produire 6598 conformations. Puis, un vecteur de traces a été extrait qui décrit chaque conformation. Cette relation plusieurs-a-un entre vecteurs de traces et les molécules s'appelle le problème de multi-instance. Lors de l'apprentissage d'un classificateur pour ces données, la classification doit classer une molécule comme musc Si l'un de ses conformations est classé comme un musc. Une molécule doit être classée comme non-musc si aucun de ses conformations n'est classé comme un musc (Dietterich et al., 1994).

B.1.15 Penbased

une base de traces de chiffres faite par la collecte de 250 échantillons provenant de 44 auteurs, en utilisant seulement les coordonnées bidimensionnelles, l'information représentée comme vecteurs de traces de longueur constante, qui ont été redimensionnés pour huit points par chiffres (donc l'ensemble de données contient 8 points coordonnées 16 attributs)[12]. L'étiquette de classe représente le code du chiffre écrit (Alimoglu et al., 1996).

B.1.16 Sonar

Cet ensemble de traces contient des signaux obtenus à partir d'une variété de différents angles d'aspect, s'étendant sur 90 degrés pour les mines et 180 degrés pour les roches[13]. Chaque motif est un ensemble de 60 numéros dont l'intervalle [0.0-1.0], où chaque nombre représente l'énergie dans une bande de fréquence particulière, intégré sur une certaine période de temps. L'attribut de sortie contient la lettre R si l'objet est une roche et M se il est une mine (cylindre métallique) (Gorman and Sejnowski, 1988).

B.1.17 SpamBase

cette base de traces contient des informations sur 4597 messages e-mail. La tâche consiste à déterminer si un e-mail donnée est un spam (classe 1) ou non (classe 2), en fonction de son contenu (Jones et al., 1990)[14]. La plupart des attributs indiqué si un mot ou caractère particulier était souvent produit dans l'email. Voici les définitions des attributs:

- 48 attributs réels continues de Type word_freq_"WORD" = pourcentage de mots dans l'email qui correspondent "WORD". Un "mot" dans ce cas est une chaîne de caractères alphanumériques

[11] http://sci2s.ugr.es/keel/dataset.php?cod=174
[12] http://sci2s.ugr.es/keel/dataset.php?cod=70
[13] http://sci2s.ugr.es/keel/dataset.php?cod=85
[14] http://sci2s.ugr.es/keel/dataset.php?cod=109

délimités par des caractères non-alphanumériques ou de fin de chaîne.

- 6 attributs réels continues de Type char_freq_"CHAR" = pourcentage de caractères dans l'e-mail qui correspondent.
- Un véritable attribut continu de Type Capital_run_length_average = longueur moyenne des séquences ininterrompues de lettres majuscules.
- 1 attribut entier continue de Type Capital_run_length_longest = longueur de la séquence la plus longue ininterrompues de lettres majuscules.
- 1 attribut entier continue de Type Capital_run_length_total = nombre total de lettres majuscules dans l'email.

B.1.18 SpectHeart

la base de traces des patients décrit le diagnostic des maladies Cardiaque utilisant l'image Single Proton Emission Computed Tomography (SPECT)[15]. Chaque patient est classé soit en cas normal (0) ou anormal (1) (Kurgan et al., 2001).

B.1.19 Thyiroid

Cet ensemble de traces des patients de la maladie thyroïdienne est l'un des plusieurs bases de traces sur la thyroïde disponibles aussi dans la base UCI [16]. La tâche consiste à détecter si un patient donné est normal (1) ou souffre d'hyperthyroïdie (2) ou l'hypothyroïdie (3) (Quinlan et al., 1987a).

B.1.20 Tiger

Cette base consiste à identifier un ensemble d'objets cibles dans une image. La difficulté dans ce genre de problème réside dans le fait que chaque image plusieurs objets hétérogènes[17]. Ainsi, se baser sur la description globale de l'image est assez large pour atteindre une bonne classification. Même si les images pertinentes sont fournies, identifier quel objet(s) dans les exemples d'images sont pertinentes reste un problème difficile dans le cadre de l'apprentissage supervisé. Toutefois, chaque image peut être traitée comme un sac de segments qui sont modélisés comme des traces, et le point concept représentant l'objet cible. Cet ensemble de traces considère ensembles de traces représentant des tigres. Chaque ensemble de traces comprend 100 images qui contiennent les tigres et les 100 autres images qui contiennent d'autres différents animaux. L'objectif final consiste à distinguer contenant les tigres de ceux qui n'en contiennent pas.

[15] http://sci2s.ugr.es/keel/dataset.php?cod=185
[16] http://sci2s.ugr.es/keel/dataset.php?cod=67
[17] http://sci2s.ugr.es/keel/dataset.php?cod=175

B.1.21 Vehicle

le but est de classer une silhouette donné que l'un des quatre types de véhicules, en utilisant un ensemble de traces extraites à partir de la silhouette[18]. Le véhicule peut être vu de l'un des nombreux angles différents. Les traces ont été extraites des silhouettes par le HIPS (Hierarchical Image Processing System) de BINATTS. Les images ont été acquises en fixant une caméra regardant vers le bas sur le modèle de véhicule à partir d'un angle d'élévation fixe (34,2 degrés à l'horizontale). Les véhicules ont été placés sur une surface rétroéclairé diffuse (visionneuse). Les véhicules ont été peints en noir mate afin de minimiser les faits saillants. Les images ont été capturées en utilisant un SIR 4000 Framestore connecté à un VAX 750. Toutes les images ont été capturées avec une résolution spatiale de 128x128 pixels quantifiés à 64 niveaux de gris. Ces images ont été seuillés pour produire des silhouettes de véhicules binaires, niée (de se conformer aux exigences de traitement de BINATTS) et par la suite soumis à shrink-expand-expand-shrink HIPS modules pour supprimer le bruit de l'image (Siebert, 1987).

B.1.22 Vowel

Vowel est une base de traces pour la reconnaissance des voyelles de l'anglais britanniques des différents speakers. Le problème s'agit d'un arrangement tridimensionnel[19]: données de voyelles [haut-parleur, voyelle, entrée]. Les haut-parleurs sont indexés par des entiers de 0 à 89. (En fait, il y a quinze haut-parleurs individuels, chacun disant chaque voyelle six fois.) Les voyelles sont indexées par des entiers de 0 à 10. Pour chaque énoncé, il y a dix valeurs d'entrée à virgule flottante, avec des indices de tableaux 0-9. Le problème est de former le réseau aussi bien que possible en utilisant uniquement les traces "haut-parleurs" 0-47, puis pour tester le réseau sur des speakers de 48 à 89 (Deterding, 1990).

[18] http://sci2s.ugr.es/keel/dataset.php?cod=68
[19] http://sci2s.ugr.es/keel/dataset.php?cod=113

Références

Abbasi, M. and Kashiyarndi, S. (2006). Clinical decision support systems: A discussion on different methodologies used in health care. Marlaedalen University Sweden. Available at: http://www. idt. mdh. se/kurser/ct3340/ht10/FinalPapers/15-Abbasi Kashiyarndi. pdf.[Accessed on: 25 Fev 2014].

Abidin, B., Dom, R. M., Rahman, A. R. A., Bakar, R. A., Demiralp, M., Baykara, N., and Mastorakis, N. (2009). Use of fuzzy neural network to predict coronary heart disease in a malaysian sample. In WSEAS International Conference. Proceedings. Mathematics and Computers in Science and Engineering, number 8. World Scientific and Engineering Academy and Society.

Aha, D. W. (1997). Editorial. In Lazy Learning, pages 7–10. Springer.

Alcock, R. J., Manolopoulos, Y., et al. (1999). Time-series similarity queries employing a feature-based approach. In 7th Hellenic conference on informatics, Ioannina, Greece, pages 1–9.

Alemdar, N. M. and Özyildirim, S. (1998). A genetic game of trade, growth and externalities. Journal of Economic Dynamics and Control, 22(6):811–832.

Alimoglu, F., Doc, D., Alpaydin, E., and Denizhan, Y. (1996). Combining multiple classifiers for pen-based handwritten digit recognition.

Amalberti, R. (1996). La conduite de systèmes à risques: le travail à l'hôpital. Presses universitaires de France.

Anderson, J. R. (2013). The architecture of cognition. Psychology Press.

Anooj, P. (2012). Clinical decision support system: risk level prediction of heart disease using decision tree fuzzy rules. Int J Res Rev Comput Sci, 3(3):1659–1667.

Asuncion, A. and Newman, D. (2007). Uci machine learning repository.

Axelrod, R. (1987). The evolution of strategies in the iterated prisoner's dilemma. The dynamics of norms, pages 1–16.

Bachimont, B. (2004a). Arts et sciences du numérique: ingénierie des connaissances et critique de la raison computationnelle. Mémoire de HDR.

Bachimont, B. (2004b). Pourquoi n'y a-t-il pas d'expérience en ingénierie des connaissances? In 15èmes Journées francophones d'ingénierie des Connaissances, pages 53–64.

Badre, A. and Santos, P. J. (1991). A knowledge-based system for capturing human-computer interaction events: Chime.

Bain, M. and Muggleton, S. (1994). Learning optimal chess strategies. In Machine intelligence 13, pages 291–309. Oxford University Press, Inc.

Banet, A. (2010). Conscience du risque et attitudes face aux risques chez les motocyclistes. PhD thesis, Lyon 2.

Bashir, S., Qamar, U., and Khan, F. H. (2015). Bagmoov: A novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting. Australasian Physical & Engineering Sciences in Medicine, pages 1–19.

Bates, P. C. (1995). Debugging heterogeneous distributed systems using event-based models of behavior. ACM Transactions on Computer Systems (TOCS), 13(1):1–31.

Beauvisage, T. (2004). Sémantique des parcours des utilisateurs sur le web. PhD in Sciences du Langage, University of Paris X, Nanterre.

Bekkerman, R., El-Yaniv, R., Tishby, N., and Winter, Y. (2003). Distributional word clusters vs. words for text categorization. The Journal of Machine Learning Research, 3:1183–1208.

Ben Ishak, A. (2007). Séléction de variables par les machines à vecteurs supports pour la discrimination binaire et multiclasse en grande dimension. PhD thesis, Aix Marseille 2.

Bermejo, P., de la Ossa, L., Gámez, J. A., and Puerta, J. M. (2012). Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking. Knowledge-Based Systems, 25(1):35–44.

Bhatla, N. and Jyoti, K. (2012). A novel approach for heart disease diagnosis using data mining and fuzzy logic. International Journal of Computer Applications, 54(17):16–21.

Blackburn, H., KEYS, A., SIMONSON, E., RAUTAHARJU, P., and PUNSAR, S. (1960). The electrocardiogram in population studies a classification system. Circulation, 21(6):1160–1175.

Blake, C. and Merz, C. J. (1998). {UCI} repository of machine learning databases.

Blansché, A. (2006). Classification non supervisée avec pondération d'attributs par des méthodes évolutionnaires. PhD thesis, Strasbourg 1.

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. Knowledge and information systems, 34(3):483–519.

Bouroche, J. and Saporta, G. (1992). L'analyse des donnÃ© e. Paris, Presses Universitaires de France.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). Classification and regression trees. CRC press.

Brinkman, W.-P., Gray, P., and Renaud, K. (2006). Computer-assisted recording, pre-processing and analysis of user interaction data. In Proc. HCI, volume 2006, pages 273–275.

Burton, A. N. and Kelly, P. H. (2006). Performance prediction of paging workloads using lightweight tracing. Future Generation Computer Systems, 22(7):784–793.

Camhi, E. (2004). Monitoring system. US Patent 6,762,684.

Catledge, L. D. and Pitkow, J. E. (1995). Characterizing browsing strategies in the world-wide web. Computer Networks and ISDN systems, 27(6):1065–1073.

Charbonnaud C, Mercier, R. (2005). Logmanager : manage your security logs.

Chauraisa, V. and Pal, S. (2013). Early prediction of heart diseases using data mining techniques. Carib. j. SciTech, 1:208–217.

Chen, A. H., Huang, S.-Y., Hong, P.-S., Cheng, C.-H., and Lin, E.-J. (2011). Hdps: heart disease prediction system. In Computing in Cardiology, 2011, pages 557–560. IEEE.

Chen, J., Sun, L., Zaïane, O. R., and Goebel, R. (2004). Visualizing and discovering web navigational patterns. In Proceedings of the 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS 2004, pages 13–18. ACM.

Chi, E. H. (2002). Improving web usability through visualization. Internet Computing, IEEE, 6(2):64–71.

Cios, K. J. and Moore, G. W. (2002). Uniqueness of medical data mining. Artificial intelligence in medicine, 26(1):1–24.

Cockburn, A., McKenzie, B., and JasonSmith, M. (2002). Pushing back: evaluating a new behaviour for the back and forward buttons in web browsers. International Journal of Human-Computer Studies, 57(5):397–414.

Coste-Manière, È., Adhami, L., Mourgues, F., and Carpentier, A. (2003). Planning, simulation, and augmented reality for robotic cardiac procedures: the stars system of the chir team. In Seminars in thoracic and cardiovascular surgery, volume 15, pages 141–156. Elsevier.

Cover, T. M. (1991). Ja thomas elements of information theory.

Crochet, D., Lefevre, M., Grossetete, R., Bouhour, J., Helias, J., GHIDALIA, S., and Delumeau, J. (1990). Evaluation comparée de l'irm, de l'échocardiographie et du cathétérisme pour le diagnostic des cardiopathies congénitales. Archives des maladies du coeur et des vaisseaux, 83(5):681–686.

Das, R., Turkoglu, I., and Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. Expert systems with applications, 36(4):7675–7680.

Dash, M. and Liu, H. (1997). Feature selection for classification. Intelligent data analysis, 1(1):131–156.

Dash, M. and Liu, H. (2003). Consistency-based search in feature selection. Artificial intelligence, 151(1):155–176.

Dash, M., Liu, H., and Motoda, H. (2000). Consistency based feature selection. In Knowledge Discovery and Data Mining. Current Issues and New Applications, pages 98–109. Springer.

Dash, M., Liu, H., and Yao, J. (1997). Dimensionality reduction of unsupervised data. In Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on, pages 532–539. IEEE.

Davis, L. (1987). Genetic algorithms and simulated annealing.

Dawid, H. (2011). Adaptive learning by genetic algorithms: Analytical results and applications to economic models. Springer Science & Business Media.

DeBusk, R. F., Miller, N. H., Superko, H. R., Dennis, C. A., Thomas, R. J., Lew, H. T., Berger, W. E., Heller, R. S., Rompf, J., Gee, D., et al. (1994). A case-management system for coronary risk factor modification after acute myocardial infarction. Annals of Internal Medicine, 120(9):721–729.

Dechter, R. and Pearl, J. (1985). Generalized best-first search strategies and the optimality of a*. Journal of the ACM (JACM), 32(3):505–536.

Degoulet, P. and Fieschi, M. (1998). Informatique médicale. Elsevier Masson.

Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. The annals of mathematical statistics, pages 325–339.

Derrac, J., GARCia, S., Sanchez, L., and Herrera, F. (2015). Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework.

Deterding, D. H. (1990). Speaker normalisation for automatic speech recognition. PhD thesis, University of Cambridge.

Dias, D. B., Madeo, R. C., Rocha, T., Bíscaro, H. H., and Peres, S. M. (2009). Hand movement recognition for brazilian sign language: a study using distance-based neural networks. In Neural Networks, 2009. IJCNN 2009. International Joint Conference on, pages 697–704. IEEE.

Diday, E. (1982). Eléments d'analyse de données. Bordas Editions.

Diekert, V., Rozenberg, G., and Rozenburg, G. (1995). The book of traces, volume 15. World Scientific.

Dietterich, T. G., Jain, A. N., Lathrop, R. H., and Lozano-Perez, T. (1994). A comparison of dynamic reposing and tangent distance for drug activity prediction. Advances in Neural Information Processing Systems, pages 216–216.

Dipchand, A., Bharat, W., Manlhiot, C., Safi, M., Lobach, N., and McCrindle, B. (2007). 512: Dobutamine stress echocardiography for the assessment of coronary artery disease in paediatric heart transplant recipients. The Journal of Heart and Lung Transplantation, 26(2):S244.

Dubois, J., Dao-Duy, J., and Eldika, S. (2000). L'analyse des traces informatiques des usages: un outil pour valider la conception d'un site web. Actes des rencontres jeunes chercheurs en Interaction Homme-Machine 2000, pages 85–89.

Ducasse, S., Girba, T., and Wuyts, R. (2006). Object-oriented legacy system trace-based logic testing. In Software Maintenance and Reengineering, 2006. CSMR 2006. Proceedings of the 10th European Conference on, pages 10–pp. IEEE.

Etiévant, H. (2004). Journalisation des évènements avec l'api logging de java.

Evans, B. and Fisher, D. (1994). Overcoming process delays with decision tree induction. IEEE expert, 9(1):60–66.

Falzon, P. (1991). Les activités verbales dans le travail. op. cit.[29], pages 229–250.

Féraud, R., Boullé, M., Clérot, F., Fessant, F., and Lemaire, V. (2010). The orange customer analysis platform. In Advances in Data Mining. Applications and Theoretical Aspects, pages 584–594. Springer.

Fidele, B., Cheeneebash, J., Gopaul, A., and Goorah, S. S. (2009). Artificial neural network as a clinical decision-supporting tool to predict cardiovascular disease. Trends in Applied Sciences Research, 4(1):36–46.

Frey, P. W. and Slate, D. J. (1991). Letter recognition using holland-style adaptive classifiers. Machine learning, 6(2):161–182.

Galan, J.-P. (2002). L'analyse des fichiers log pour étudier l'impact de la musique sur le comportement des visiteurs d'un site web culturel. Actes du 18 ème Congrès International de l'Association Française du Marketing (AFM).

Georgeon, O. (2008). Analyse de traces d'activité pour la modélisation cognitive: application à la conduite automobile [pdf].

Gibler, W. B., Cannon, C. P., Blomkalns, A. L., Char, D. M., Drew, B. J., Hollander, J. E., Jaffe, A. S., Jesse, R. L., Newby, L. K., Ohman, E. M., et al. (2005). Practical implementation of the guidelines for unstable angina/non–st-segment elevation myocardial infarction in the emergency department a scientific statement from the american heart association council on clinical cardiology (subcommittee on acute cardiac care), council on cardiovascular nursing, and quality of care and outcomes research interdisciplinary working group, in collaboration with the society of chest pain centers. Circulation, 111(20):2699–2710.

Gillet, B. (1987). Le psychologue et l'ergonomie. Etablissements d'applications psychotechniques.

Golberg, D. E. (1989). Genetic algorithms in search, optimization, and machine learning. Addion wesley, 1989.

Goldberg, D. E., Corruble, V., Ganascia, J.-G., and Holland, J. (1994). Algorithmes génétiques: exploration, optimisation et apprentissage automatique. Addison-Wesley France.

Gorman, R. P. and Sejnowski, T. J. (1988). Analysis of hidden units in a layered network trained to classify sonar targets. Neural networks, 1(1):75–89.

Gray, P., McLeod, I., Draper, S., Crease, M., and Thomas, R. (2004). A distributed usage monitoring system.

Greensmith, J. (2003). New frontiers for an artificial immune system. Digital Media Systems Laboratory HP Laboratories Bristol HPL-2003-204, 45.

Guérif, S. (2006). Réduction de dimension en apprentissage numérique non supervisé. PhD thesis, Paris 13.

Guimbretière, F., Dixon, M., and Hinckley, K. (2007). Experiscope: an analysis tool for interaction data. In Proceedings of the SIGCHI conference on Human factors in computing systems, pages 1333–1342. ACM.

Guo, Y., Wang, Y., Kong, D., and Shu, X. (2014). Automatic classification of intracardiac tumor and thrombi in echocardiography based on sparse representation.

Güvenir, H., Acar, B., Demiröz, G., et al. (1997). A supervised machine learning algorithm for arrhythmia analysis. In Computers in Cardiology 1997, pages 433–436. IEEE.

Güvenir, H. A., Demiröz, G., and Ilter, N. (1998). Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals. Artificial intelligence in medicine, 13(3):147–165.

Guyon, I., Aliferis, C., and Elisseeff, A. (2007). Causal feature selection. Computational methods of feature selection, pages 63–86.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. ACM SIGKDD explorations newsletter, 11(1):10–18.

Halvey, M., Keane, M. T., and Smyth, B. (2005). Time-based segmentation of log data for user navigation prediction in personalization. In Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on, pages 636–640. IEEE.

Hanoune, M. and Benabbou, F. (2006). Traitement et exploration du fichier log du serveur web, pour l'extraction des connaissances: Web usage mining. Afrique Science: Revue Internationale des Sciences et Technologie, 2(3).

Hawkey, K. and Inkpen, K. M. (2005). Privacy gradients: exploring ways to manage incidental information during co-located collaboration. In CHI'05 Extended Abstracts on Human Factors in Computing Systems, pages 1431–1434. ACM.

Haykin, S. (1999). Adaptive filters. Signal Processing Magazine, 6.

Hedeshi, N. G. and Abadeh, M. S. (2014). Coronary artery disease detection using a fuzzy-boosting pso approach. Computational intelligence and neuroscience, 2014:6.

Hendryx, M. and Zullig, K. J. (2009). Higher coronary heart disease and heart attack morbidity in appalachian coal mining regions. Preventive Medicine, 49(5):355–359.

Hétier, M. (2009). Analyse et quantification des comportements des conducteurs automobiles lors des phases de pré-crash: Contribution au développement d'un modèle de détection des postures de conduite en temps réel. PhD thesis, Valenciennes.

Hilbert, D. M. and Redmiles, D. F. (2000). Extracting usability information from user interface events. ACM Computing Surveys (CSUR), 32(4):384–421.

Holland, J. H. (1992). Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press.

Hong, R., Wang, M., Gao, Y., Tao, D., Li, X., and Wu, X. (2014). Image annotation by multiple-instance learning with discriminative feature mapping and selection. Cybernetics, IEEE Transactions on, 44(5):669–680.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of educational psychology, 24(6):417.

Iantovics, B., Marusteri, M., Kountchev, R., Zamfirescu, C.-B., and Crainicu, B. Intelligent cmds medical agents with learning capacity.

Jang, J.-S. R., Sun, C.-T., and Mizutani, E. (1997). Neuro-fuzzy and soft computing; a computational approach to learning and machine intelligence.

Jansen, B. J. and Pooch, U. (2001). A review of web searching studies and a framework for future research. Journal of the American Society for Information science and Technology, 52(3):235–246.

Jilani, T. A., Yasin, H., Yasin, M., and Ardil, C. (2009). Acute coronary syndrome prediction using data mining techniques: an application. World Academy of Science, Engineering and Technology, 59(4):295–299.

Jin, C., Jin, S.-W., and Qin, L.-N. (2012). Attribute selection method based on a hybrid bpnn and pso algorithms. Applied Soft Computing, 12(8):2147–2155.

Jones, R. D., Lee, Y., Barnes, C., Flake, G., Lee, K., Lewis, P., and Qian, S. (1990). Function approximation and time series prediction with neural networks. In Neural Networks, 1990., 1990 IJCNN International Joint Conference on, pages 649–665. IEEE.

Jong, K. D. (1980). Adaptive system design: a genetic approach. Systems, Man and Cybernetics, IEEE Transactions on, 10(9):566–574.

Kabir, M. M., Shahjahan, M., and Murase, K. (2011). A new local search based hybrid genetic algorithm for feature selection. Neurocomputing, 74(17):2914–2928.

Kahramanli, H. and Allahverdi, N. (2008). Design of a hybrid system for the diabetes and heart diseases. Expert Systems with Applications, 35(1):82–89.

Kamin, S. (1990). A debugging environment for functional programming in centaur.

Karaolis, M., Moutiris, J. A., and Pattichis, C. S. (2008). Assessment of the risk of coronary heart event based on data mining. In BioInformatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference on, pages 1–5. IEEE.

Karoli, R., Fatima, J., Singh, P., and Kazmi, K. I. (2012). Acute myocardial involvement after heroin inhalation. Journal of pharmacology & pharmacotherapeutics, 3(3):282.

Kaushal, R., Shojania, K. G., and Bates, D. W. (2003). Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. Archives of internal medicine, 163(12):1409–1416.

Kerr, W. T., Lau, E. P., Owens, G. E., and Trefler, A. (2012). The future of medical diagnostics: large digitized databases. The Yale journal of biology and medicine, 85(3):363.

Khatibi, V. and Montazer, G. A. (2010). A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment. Expert Systems with Applications, 37(12):8536–8542.

Kibler, D., Aha, D. W., and Albert, M. K. (1989). Instance-based prediction of real-valued attributes. Computational Intelligence, 5(2):51–57.

Kim, Y.-H., Gitelman, D. R., Nobre, A. C., Parrish, T. B., LaBar, K. S., and Mesulam, M.-M. (1999). The large-scale neural network for spatial attention displays multifunctional overlap but differential asymmetry. Neuroimage, 9(3):269–277.

Klir, G. and Yuan, B. (1995). Fuzzy sets and fuzzy logic, volume 4. Prentice Hall New Jersey.

Kochurani, O., Aji, S., and Kaimal, M. (2007). A neuro fuzzy decision tree model for predicting the risk in coronary artery disease. In Intelligent Control, 2007. ISIC 2007. IEEE 22nd International Symposium on, pages 166–171. IEEE.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. Artificial intelligence, 97(1):273–324.

Koller, D. and Sahami, M. (1996). Toward optimal feature selection.

Korfiatis, V. C., Asvestas, P. A., Delibasis, K. K., and Matsopoulos, G. K. (2013). A classification system based on a new wrapper feature selection algorithm for the diagnosis of primary and secondary polycythemia. Computers in biology and medicine, 43(12):2118–2126.

Krishnakumar, K. and Goldberg, D. E. (1992). Control system optimization using genetic algorithms. Journal of Guidance, Control, and Dynamics, 15(3):735–740.

Kudo, M. and Sklansky, J. (2000). Comparison of algorithms that select features for pattern classifiers. Pattern recognition, 33(1):25–41.

Kurgan, L. A., Cios, K. J., Tadeusiewicz, R., Ogiela, M., and Goodenday, L. S. (2001). Knowledge discovery approach to automated cardiac spect diagnosis. Artificial intelligence in medicine, 23(2):149–169.

Kurt, I., Ture, M., and Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Systems with Applications, 34(1):366–374.

Laflaquière, J. (2009). Conception de système à base de traces numériques pour les environnements informatiques documentaires. PhD thesis, Université de Technologie de Troyes.

Langlotz, C. P. and Shortliffe, E. H. (1983). Adapting a consultation system to critique user plans. International Journal of Man-Machine Studies, 19(5):479–496.

Larousse, P. (2013). Le grand dictionnaire universel du xixe siècle, édition.

Lee, B. (1996). Remote diagnostics and product lifecycle monitoring for high-end appliances: a new internet-based approach utilizing intelligent software agents. In Proceedings of the Appliance Manufacturer Conference.

Leont'ev, A. N., Dupond, G., and Molinier, G. (1984). Activité, conscience, personnalité. Editions du progrès.

Leplat, J. (2003). 1. la modélisation en ergonomie à travers son histoire. Le Travail humain, 1:1–26.

Lerman, I.-C. and Ngouenet, R. (1995). Algorithmes génétiques séquentiels et parallèles pour une représentation affine des proximités.

Lewis, D. D. (1998). Naive (bayes) at forty: The independence assumption in information retrieval. In Machine learning: ECML-98, pages 4–15. Springer.

Lin, C.-T. and Lee, C. G. (1991). Neural-network-based fuzzy logic control and decision system. Computers, IEEE Transactions on, 40(12):1320–1336.

Loève, M. and Lévy, P. (1948). Processus stochastiques et mouvement Brownien.

Malergue, M., Temkine, J., Slama, M., Dibie, A., Ledavay, M., BENRABBHA, T., LABORDE, F., and LECOMPTE, Y. (1992). Intérêts de l'échocardiographie tranœsophagienne systématique postopératoire précoce des remplacements valvulaires mitraux: une étude prospective sur 50 patients. Archives des maladies du coeur et des vaisseaux, 85(9):1299–1304.

Mamdani, E. H. and Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. International journal of man-machine studies, 7(1):1–13.

Marateb, H. R. and Goudarzi, S. (2015). A noninvasive method for coronary artery diseases diagnosis using a clinically-interpretable fuzzy rule-based system. Journal of Research in Medical Sciences, 20(3).

Marco, N., Godart, C., Désidéri, J.-A., Mantel, B., and Périaux, J. (1996). A genetic algorithm compared with a gradient-based method for the solution of an active-control model problem.

Mejía-Lavalle, M., Sucar, E., and Arroyo, G. (2006). Feature selection with a perceptron neural net. In Proceedings of the international workshop on feature selection for data mining, pages 131–135.

Michalewicz, Z., Janikow, C. Z., and Krawczyk, J. B. (1992). A modified genetic algorithm for optimal control problems. Computers & Mathematics with Applications, 23(12):83–94.

Michalski, R. S. and Chilausky, R. L. (1980). Learning by being told and learning from examples: An experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis. International Journal of Policy Analysis and Information Systems, 4(2):125–161.

Mohr, F. W., Falk, V., Diegeler, A., Walther, T., Gummert, J. F., Bucerius, J., Jacobs, S., and Autschbach, R. (2001). Computer-enhanced "robotic" cardiac surgery: experience in 148 patients. The Journal of thoracic and cardiovascular surgery, 121(5):842–853.

Mokeddem, S. and Atmani, B. (2016). Assessment of clinical decision support systems for predicting coronary heart disease. international journal of Operations Research and Information Systems (IJORIS), 7(3).

Mokeddem, S., Atmani, B., and Mokaddem, M. (2014). A new approach for coronary artery diseases diagnosis based on genetic algorithm. International Journal of Decision Support System Technology (IJDSST), 6(4):1–15.

Mokeddem, S., Atmani, B., and Mokaddem, M. (2016). An effective feature selection approach driven genetic algorithm wrapped bayes naïve. International Journal of Decision Support System Technology (IJDSST), 8(3).

Nahar, J., Imam, T., Tickle, K. S., and Chen, Y.-P. P. (2013). Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Applications, 40(4):1086–1093.

Nakao, S., Kuwano, T., Tsutsumi-Miyahara, C., Ueda, S.-i., Kimura, Y. N., Hamano, S., Sonoda, K.-h., Saijo, Y., Nukiwa, T., Strieter, R. M., et al. (2005). Infiltration of cox-2–expressing macrophages is a prerequisite for il-1β–induced neovascularization and tumor growth. Journal of Clinical Investigation, 115(11):2979.

Nalinipriya, G., Kannan, A., and Anandhakumar, P. (2012). Performance analysis of classifiers for multivariate coronary artery disease dataset using renowned metrics. European Journal of Scientific Research, 86(4):565–572.

Neal, A. S. and Simons, R. M. (1984). Playback: A method for evaluating the usability of software and its documentation. IBM Systems Journal, 23(1):82–96.

Neisser, U. (1976). Cognition and reality: Principles and implications of cognitive psychology. WH Freeman/Times Books/Henry Holt & Co.

Nichol, G., Thomas, E., Callaway, C. W., Hedges, J., Powell, J. L., Aufderheide, T. P., Rea, T., Lowe, R., Brown, T., Dreyer, J., et al. (2008). Regional variation in out-of-hospital cardiac arrest incidence and outcome. Jama, 300(12):1423–1431.

Nielsen, J. (1994). Usability engineering. Elsevier.

Nienaber, C. A., von Kodolitsch, Y., Nicolas, V., Siglow, V., Piepho, A., Brockhoff, C., Koschyk, D. H., and Spielmann, R. P. (1993). The diagnosis of thoracic aortic dissection by noninvasive imaging procedures. New England Journal of Medicine, 328(1):1–9.

Niharika, S., Latha, V. S., and Lavanya, D. (2012). A survey on text categorization. International Journal of Computer Trends and Technologyvolume3Issue1-2012.

Noordewier, M. O., Towell, G. G., and Shavlik, J. W. (1991). Training knowledge-based neural networks to recognize genes in dna sequences. In Advances in neural information processing systems, pages 530–536.

of Cardiology, A. C., Association, A. H., et al. (1991). Guidelines and Indications for Coronary Artery Bypass Graft Surgery: A Report of the American College of Cardiology/American Heart; Association Task Force on Assessment of Diagnostic And; Therapeutic Cardiovascular Procedures. Md.

Olsen, D. R. and Halversen, B. W. (1988). Interface usage measurements in a user interface management system. In Proceedings of the 1st annual ACM SIGGRAPH symposium on User Interface Software, pages 102–108. ACM.

Oreski, S. and Oreski, G. (2014). Genetic algorithm-based heuristic for feature selection in credit risk assessment. Expert systems with applications, 41(4):2052–2064.

Osheroff, J. A., Teich, J. M., Middleton, B., Steen, E. B., Wright, A., and Detmer, D. E. (2007). A roadmap for national action on clinical decision support. Journal of the American medical informatics association, 14(2):141–145.

Ottogalli, F.-G. and Vincent, J.-M. (1999). Mise en cohérence et analyse de traces logicielles multiniveaux. Calculateurs parallèles, 11(2):211–227.

Özyildirim, S. (1996). Three-country trade relations: A discrete dynamic game approach. Computers & Mathematics with Applications, 32(5):43–56.

Özyildirim, S. (1997). Computing open-loop noncooperative solution in discrete dynamic games. Journal of Evolutionary Economics, 7(1):23–40.

Palaniappan, S. and Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. In Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on, pages 108–115. IEEE.

Patarin, S. (1999). Pandora: un système de collecte de traces du trafic web de communautés d'utilisateurs réparties.

Patil, S. B. and Kumaraswamy, Y. (2009). Intelligent and effective heart attack prediction system using data mining and artificial neural network. European Journal of Scientific Research, 31(4):642–656.

Pereira, R. (2000). Genetic algorithm optimisation for finance and investments.

Pietra, S. D., Pietra, V. D., and Lafferty, J. (1997). Inducing features of random fields. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 19(4):380–393.

Pike, N. A. and Gundry, S. R. (2003). Robotically assisted cardiac surgery: minimally invasive techniques to totally endoscopic heart surgery. Journal of Cardiovascular Nursing, 18(5):382–388.

Pirolli, P., Fu, W.-T., Reeder, R., and Card, S. K. (2002). A user-tracing architecture for modeling interaction with the world wide web. In Proceedings of the Working Conference on Advanced Visual Interfaces, pages 75–83. ACM.

Platt, J. et al. (1999). Fast training of support vector machines using sequential minimal optimization. Advances in kernel methods—support vector learning, 3.

Polat, K., Güneş, S., and Tosun, S. (2006). Diagnosis of heart disease using artificial immune recognition system and fuzzy weighted pre-processing. Pattern Recognition, 39(11):2186–2193.

Prié, A. M.-Y. (2006). Une théorie de la trace informatique pour faciliter l'adaptation dans la confrontation logique d'utilisation/logique de conception.

Proulx-Sammut, L. (1998). La ménopause mieux comprise. Une véritable amie, XIV, 8:1–3.

Quinlan, J. R. (1996). Bagging, boosting, and c4. 5. In AAAI/IAAI, Vol. 1, pages 725–730.

Quinlan, J. R., Compton, P. J., Horn, K., and Lazarus, L. (1987a). Inductive knowledge acquisition: a case study. In Proceedings of the Second Australian Conference on Applications of expert systems, pages 137–156. Addison-Wesley Longman Publishing Co., Inc.

Quinlan, J. R., Compton, P. J., Horn, K. A., and Lazarus, L. (1987b). Inductive knowledge acquisition: A case study. In Proceedings of the Second Australian Conference on Applications of Expert Systems, pages 137–156, Boston, MA, USA. Addison-Wesley Longman Publishing Co., Inc.

Rabardel, P. and Samurçay, R. (2001). From artifact to instrument-mediated learning. In Symposium on New challenges to research on Learning, pages 21–23.

Robnik-Šikonja, M. and Kononenko, I. (2003). Theoretical and empirical analysis of relieff and rrelieff. Machine learning, 53(1-2):23–69.

Rossi, F., Lechevallier, Y., El Golli, A., and AxIS, P. (2005). Visualisation de la perception d'un site web par ses utilisateurs. In EGC, pages 563–574.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. Science, 290(5500):2323–2326.

Ruiz, R., Riquelme, J. C., Aguilar-Ruiz, J. S., and García-Torres, M. (2012). Fast feature selection aimed at high-dimensional data via hybrid-sequential-ranked searches. Expert Systems with Applications, 39(12):11094–11102.

Sanderson, P. M. and Fisher, C. (1994). Exploratory sequential data analysis: Foundations. Human–Computer Interaction, 9(3-4):251–317.

Saporta, G. (2011). Probabilités, analyse des données et statistique. Editions Technip.

Schaffernicht, E., Stephan, V., and Groß, H.-M. (2007). An efficient search strategy for feature selection using chow-liu trees. In Artificial Neural Networks–ICANN 2007, pages 190–199. Springer.

Schlimmer, J. C. (1987). Concept acquisition through representational adjustment.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In Computational learning theory, pages 416–426. Springer.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. Neural computation, 10(5):1299–1319.

Sémani-Delmi, D. (2004). Une méthode supervisée de sélection et de discrimination avec rejet: application au projet Aqu@ thèque. PhD thesis, La Rochelle.

Setiawan, N. A., Venkatachalam, P., and Hani, A. F. M. (2009). Diagnosis of coronary artery disease using artificial intelligence based decision support system. In Proceedings of the International Conference on Man-Machine Systems, Batu Ferringhi, Penang, pages 11–13.

Shafer, G. et al. (1976). A mathematical theory of evidence, volume 1. Princeton university press Princeton.

Shouman, M., Turner, T., and Stocker, R. (2012). Integrating naive bayes and k-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients. Glob J Comput Sci Technol, pages 125–137.

Siebert, J. P. (1987). Vehicle recognition using rule based methods.

Siochi, A. C. and Ehrich, R. W. (1991). Computer analysis of user interfaces based on repetition in transcripts of user sessions. ACM Transactions on Information Systems (TOIS), 9(4):309–335.

Sperandio, J.-C. (2003). 2. modèles et formalismes, ou le fond et la forme. Le Travail humain, 1:27–75.

Srivastava, J., Cooley, R., Deshpande, M., and Tan, P.-N. (2000). Web usage mining: Discovery and applications of usage patterns from web data. ACM SIGKDD Explorations Newsletter, 1(2):12–23.

Stanton, N. A., Salmon, P. M., Walker, G. H., and Jenkins, D. (2009). Genotype and phenotype schemata as models of situation awareness in dynamic command and control teams. International Journal of Industrial Ergonomics, 39(3):480–489.

Sundar, N. A., Latha, P. P., and Chandra, M. R. (2012). Performance analysis of classification data mining techniques over heart disease database.

Tauscher, L. and Greenberg, S. (1997). How people revisit web pages: Empirical findings and implications for the design of history systems. International Journal of Human-Computer Studies, 47(1):97–137.

Thomas, R., Kennedy, G. E., Draper, S., Mancy, R., Crease, M., Evans, H., and Gray, P. (2003). Generic usage monitoring of programming students. In ASCILITE, volume 3.

TREE, A. C. A. (2014). A fast and accurate method for automatic coronary arterial tree extraction in angiograms. Journal of Computer Science, 10(10):2060–2076.

Tsipouras, M. G., Exarchos, T. P., Fotiadis, D. I., Kotsia, A. P., Vakalis, K. V., Naka, K. K., and Michalis, L. K. (2008). Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. Information Technology in Biomedicine, IEEE Transactions on, 12(4):447–458.

Vaisi-Raygani, A., Ghaneialvar, H., Rahimi, Z., Nomani, H., Saidi, M., Bahrehmand, F., Vaisi-Raygani, A., Tavilani, H., and Pourmotabbed, T. (2010). The angiotensin converting enzyme d allele is an independent risk factor for early onset coronary artery disease. Clinical biochemistry, 43(15):1189–1194.

Van Breukelen, M., Duin, R. P., Tax, D. M., and Den Hartog, J. (1998). Handwritten digit recognition by combined classifiers. Kybernetika, 34(4):381–386.

Van Der Putten, P. and Van Someren, M. (2004). A bias-variance analysis of a real world learning problem: The coil challenge 2000. Machine Learning, 57(1-2):177–195.

van Melle, W., Shortliffe, E. H., and Buchanan, B. G. (1984). Emycin: A knowledge engineer's tool for constructing rule-based expert systems. Rule-based expert systems: The MYCIN experiments of the Stanford Heuristic Programming Project, pages 302–313.

Vergara, J. R. and Estévez, P. A. (2014). A review of feature selection methods based on mutual information. Neural Computing and Applications, 24(1):175–186.

Vermersch, P. (1994). L'entretien d'explicitation en formation initiale et en formation continuale.

Viera, A. J., Garrett, J. M., et al. (2005). Understanding interobserver agreement: the kappa statistic. Fam Med, 37(5):360–363.

Warren, J., Beliakov, G., and Van der Zwaag, B. (2000). Fuzzy logic in clinical practice decision support systems. In System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on, pages 10–pp. IEEE.

Weinreich, H., Obendorf, H., Herder, E., and Mayer, M. (2006). Off the beaten tracks: exploring three aspects of web navigation. In Proceedings of the 15th international conference on World Wide Web, pages 133–142. ACM.

Witten, I. H. and Frank, E. (2005). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Wung, S.-F., Hickey, K. T., Taylor, J. Y., and Gallek, M. J. (2013). Cardiovascular genomics. Journal of Nursing Scholarship, 45(1):60–68.

Xu, L., Krzyżak, A., and Suen, C. Y. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. Systems, man and cybernetics, IEEE transactions on, 22(3):418–435.

Yan, H., Zheng, J., Jiang, Y., Peng, C., and Xiao, S. (2008). Selecting critical clinical features for heart diseases diagnosis with a real-coded genetic algorithm. Applied Soft Computing, 8(2):1105–1111.

Yang, C. S., Chung, J. Y., Yang, G.-y., Chhabra, S. K., and Lee, M.-J. (2000). Tea and tea polyphenols in cancer prevention. The Journal of nutrition, 130(2):472S–478S.

Yang, Y. and Liu, X. (1999). A re-examination of text categorization methods. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 42–49. ACM.

Yu, L. and Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In ICML, volume 3, pages 856–863.

Yu, L. and Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. The Journal of Machine Learning Research, 5:1205–1224.

Zadeh, L. A. (1965). Fuzzy sets. Information and control, 8(3):338–353.

Zhao, Z. and Liu, H. (2007). Searching for interacting features. In IJCAI, volume 7, pages 1156–1161.

Zhu, Z., Ong, Y.-S., and Dash, M. (2007). Wrapper–filter feature selection algorithm using a memetic framework. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 37(1):70–76.

Zighed, D. and Rakotomala, R. (1996). A method for non arborescent induction graphs. Loboratory ERIC, University of Lyon, 2.

Zighed, D. A., Auray, J.-P., Duru, G., and Lacassagne, E. A. (1992). SIPINA: Méthode et logiciel. Editions Alexandre Lacassagne.