REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEURE ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE ABOU-BEKR BELKAID – TLEMCEN

THÈSE LMD

Présentée à :

FACULTE DES SCIENCES – DEPARTEMENT D'INFORMATIQUE

Pour l'obtention du diplôme de :

DOCTORAT

Spécialité : *Informatique*
Apprentissage artificielle et ingénierie de connaissances

Par :

Mme AMGHAR Djazia
(EP HAMDAN)

Sur le thème

Conception d'un modèle médical à base de résumés linguistiques

---

Soutenue publiquement le 27/02/2019 l'université à Tlemcen devant le jury composé de :

Mme DIDI Fadoua	Professeure	Université de Tlemcen	Présidente
Mr CHIKH Med.Amine	Professeur	Université de Tlemcen	Directeur de thèse
Mr ABDERRAHIM Med. Amine	Maître de Conférences A	Université de Tlemcen	Examinateur
Mr BOUAMRANE Karim	Professeur	Université d'Oran	Examinateur
Mr ATMANI Baghdad	Professeur	Université d'Oran	Examinateur
Mr BENAMMAR Abdelkrim	Maître de Conférences A	Université de Tlemcen	Examinateur

Laboratoire Génie Biomédical GBM
BP 119, 13000 Tlemcen - Algérie

Remercîment

En toute gratitude, J'exprime ma profonde gratitude tout d'abord au bon Dieu Miséricordieux de m'avoir donne le courage, la volonté et la force pour réaliser ce modeste travail.

Je tiens à remercier tous celles et ceux qui m'ont aidé de près ou de loin à mener à bien ma recherche, tous celles ou ceux qui étaient disponibles et ont participé à rendre plus humaine, plus riche et enrichissante cette longue expérience. Mes remerciements iront aussi bien à mes enseignants qu'au cadre administratif de notre département.

Je témoigne mes sincères remerciements avec ma plus profonde gratitude à mon directeur de thèse Professeur CHIKH Mohammed Amine, qui, tout au long de mon parcours, m'a appuyé et m'a encouragé à donner le meilleur de moi-même. Je lui exprime toute ma gratitude pour avoir réussi à stimuler continuellement mes capacités et mon potentiel. Il a su être réconfortant et généreux sans jamais laisser de côté la rigueur ni la précision qui incombent à la recherche.

Mes vifs remerciements sont aussi adressés à tous les membres du jury.

Je remercie vivement Mme DIDI FEDOUA pour m'avoir fait l'honneur de présider mon jury de thèse. Mes vifs remerciements vont également à M.ABDERRAHIM, M. BOUAMRANE, M. ATMANI et M. BENAMMAR pour avoir bien voulu de prendre le temps d'examiner cette thèse.

Je tiens également à remercier Dr EL HABIB DAHO Mostafa pour ses conseils et son soutien moral et scientifique qui m'ont permis de mener à bien ce projet de thèse.

Je remercie également tous mes collègues de l'université de Tlemcen et particulièrement mes collègues de l'équipe CREDOM, Dr Sarah Bellaroussi, Mme Kissi wassila , Dr Beloufa Fayssal pour leur collaboration et présence.

Du fond du cœur je remercie infiniment toute ma famille et ma belle famille, en particulier : Mes Parents, Mon Marie, Ma sœur, mes frères et Mes deux adorables fils (Noureddine et Anes).

Dédicace

Par la grâce d'allah je dédie ce travail . . .
A ma douce Maman, ma conseillère, et amie fidèle, qui m'a assistée dans les moments difficiles, et qui a su me prêter l'attention, l'affection et l'amour qui m'ont permis d'écrire cette thèse dans de bonnes conditions. Tu représentes pour moi le symbole de la bonté par excellence. Tu n'as pas cessé de me soutenir, de m'encourager et de prier pour moi durant toutes les années de mes études, tu as toujours été présente à mes côtés pour me consoler quand il le fallait.
A mon très cher Père, les mots ne me suffisent pas pour exprimer l'attachement, l'amour l'affection et la profonde tendresse que je te porte pour ton soutien et tes encouragements. Tu as toujours répondu présent pour moi, tes actes étaient la lumière dans l'obscure des moments difficiles, et m'ont toujours donné la force de poursuivre mes études.
A mon cher mari, Said, pour ses nombreuses corrections, suggestions et encouragements. Je n'aurais jamais pu terminer ce travail sans sa compréhension, ses sacrifices et toute son aide afin de me laisser consacrer un maximum de temps pour ma thèse.
A ma chère sœur que j'adore le plus dans ce monde Soumia , merci pour ta générosité , tes leçons de moral et ton soutient.
Ainsi qu'à mes frères Hadi et Hamza , vous m'avez soutenu énormément. Que Dieu vous bénisses.

A ma belle famille et en particulier mes beaux parents, pour leur soutient et encouragements.

A mes rayon de soleil, mes adorable fils, Noureddine et Anes, qui me couvre de joie et tendresse. Ses sourires et ses câlins me remontent toujours le moral.
A toute la famille.
Merci à Vous Tous.

Résumé

D’une manière générale, la collecte des données médicales créent une grande base des données numériques. Notre travail de recherche vise à extraire des résumés linguistiques flous à partir d'une grande base de données médicale. De ce faite, nous proposons un modèle de diagnostique médical qui traite le problème de volumes de données médicales par la méthode de résumé linguistique, à base de calcul de cardinalité floue. Ce type de résumé permet de construire une base de connaissances réduite qui contient toutes les informations essentielles, pour une meilleure décision. Cette dernier est aussi utilisé pour construire un système interrogation flexible qui traite des requêtes sémantiques simples et des requêtes complexes en langage naturel à la l’aide d’une nouvelle approche, proposé dans notre thèse de recherche.

De ce fait, nous proposons aussi de développer un nouveau type de classifieur supervisé basé sur ces résumés linguistiques des données médicales. Pour réaliser cet objectif, nous utilisons le calcul de la similarité entre les différents ensembles flous de résumés linguistiques. La classe (type de maladie) du patient est identifiée par le calcul de l’intégrale de Sugeno.

Dans ce contexte, nous réalisons un nouveau modèle médical qui combine entre la réduction sémantique des données, la recherche sémantique de données et de créer un nouveau classifieur supervisé qui est basé sur les résumés linguistiques. Ainsi, la solution proposée a été validée expérimentalement sur des bases de données médicales réelles. Les résultats obtenus sont comparés à ceux de l’état de l’art où nous montrons l’efficacité de modèle de la déduction proposée.

Mots-clés: Données numériques médicales, résumé linguistique, requêtes floues, classification des données médicales, logique floue.

Table des matières

Remercîment..... I

Dédicace II

Résumé..... IV

Table des matières..... IV

Liste des figures IX

Liste des Tableaux..... X

Introduction général.....1

chapitre 1. Résumé linguistique 5

1.1 Introduction 6

1.2 Compression de données 6

1.2.1 Compression physique 6

1.2.2 Compression logique 7

1.3 Réduction sémantique de données 7

1.3.1 La réduction basée sur les méthodes statistiques 8

1.3.1.1 La réduction verticale 9

1.3.1.2 Réduction horizontale 9

1.3.1.3 Calcul d'agrégats 9

1.3.2 Approche basé sur les modèles 9

1.3.2.1 Le modèle de fouille de données 9

1.3.2.2 Les modèles basés sur les métadonnées 10

1.4 Résumé linguistique : 10

1.4.1 Définition Variable linguistique : 10

1.4.2 Théorie de l'ensemble flou 11

1.5 Résumé SaintEtiQ 13

1.5.1 Les étapes de la construction hiérarchique de résumés. 14

1.5.1.1 Réécriture des données : 14

1.5.1.2 Regroupement des données 15

1.6 Résumé de Yager 16

1.7 Résumé linguistique à base de calcul de la cardinalité floue 20

1.7.1 Principe de résumé linguistique à base de calcul de la cardinalité floue 21

1.7.2 Définition le Quantificateur flou 24

1.8 Conclusion 25

chapitre 2. Interrogation flexible des résumés linguistiques 26

2.1 Introduction 27

2.2 L'interrogation de la Base de données relationnel 27

2.2.1 Définition de la base de données : 27

2.2.2 Système d'interrogation flexible 28

2.2.2.1 Système de personnalisation 29

2.2.2.2 Requête flexible 29

2.2.3 Requête floue : 30

2.3 Système d'interrogation des résumés linguistiques 31

2.3.1 Le modèle SaintEtiQ 32

2.3.1.1 La Forme général de la requête : 33

2.3.1.2 Évaluation des requêtes : 34

2.3.2 Résumé de Yager 35

2.3.2.1 Langage d'interrogation de résumé Yager 36

2.3.3 Approche de L.Liétard 39

2.3.4 Résume linguistique à base de calcul de Cardinalité floue 40

2.3.4.1 La définition de résumé linguistique à base de calcul de la cardinalité floue 40

2.3.4.2 Quantité graduel 41

2.4 Conclusion 45

chapitre 3. La relation entre le résumé linguistique et la classification supervisée 46

3.1 Introduction 47

3.2 Extraction de connaissance à partir de données 47

3.2.1 Définition de Data Mining (Fouille de données) 47

3.2.1.1 L'estimation 48

3.2.1.2 La prédiction 48

3.2.1.3 Les règles d'association 48

3.2.1.4 La segmentation 48

3.2.1.5 Classification 48

3.3 La classification 48

3.3.1 L'apprentissage 49

3.3.1.1 L'apprentissage non supervisé : 49

3.3.1.2 L'apprentissage semi-supervisé (par renforcement) : 49

3.3.1.3 L'apprentissage supervisé : 49

3.3.2 Principe de la classification 50

3.3.2.1 Définition d'une classe 50

3.3.2.2 Définition d'un classifieur 50

3.3.3 Classification supervisée 51

3.3.3.1 Risque réel 52

3.3.3.2 Risque empirique 52

3.4 Les méthodes de classification 52

3.4.1 Séparateurs à Vaste Marge 52

3.4.1.1 Principe 53

3.4.1.2 Ajustement 54

3.4.1.3 Avantages et inconvénients 54

3.4.2 Arbre de décision 54

3.4.2.1 Principale de l'arbre de décision 54
3.4.2.2 Avantages et inconvénients 55
3.4.3 Réseaux de neurone 55
3.4.3.1 Principe 56
3.4.3.2 Mise en œuvre 57
3.4.3.3 Avantages et Inconvénients 57
3.4.4 Les plus proches voisins 57
3.4.4.1 Les distances 58
3.4.4.2 Avantages et inconvénients 59
3.5 Prédiction et le calcul de la distance métrique des résumés linguistique 60
3.5.1 Présentation de résumé linguistique 60
3.5.1.1 Les protoformes classiques : 60
3.5.1.2 L'extension de protoforme de Yager 62
3.5.2 Evaluation de la similarité d'une protoforme classique : 63
3.5.3 Similarité entre les ensembles des résumes linguistiques 67
3.5.3.1 Agrégation en utilisant opérateurs OWA : 68
3.5.3.2 Agrégation de Intégral Sugeno 69
3.5.4 Résumés linguistiques différentiels 70
3.5.5 Travaux de littératures 72
3.6 Conclusion 74
chapitre 4. Conception d'un modèle médical à base de résumés linguistiques 75
4.1 Introduction 76
4.2 Matériels et méthodes 76
4.2.1 Langages et outils utilisés 76
4.2.2 Les bases de données 77
4.2.2.1 PIMA 77
4.2.2.2 Wisconsin Breast Cancer (WBCD) 77
4.2.2.3 Mammographie 78
4.2.3 Critères d'évaluation 78
4.3 Architecture d'un nouveau modèle de diagnostic médical 79
- Aperçue sur l'application 80

4.4 Approche proposée des résumés linguistiques médicaux RLR-CardF 81
4.4.1 Implémentation le RLR-CardF sur une base de données médicale 84
4.4.2 Résultat et discutions 86
4.4.3 Aperçue sur l'application 88
4.5 Approche proposée d'un système d'interrogation médical flexible 90
4.5.1 Calcul le quantificateur flou par le produit scalaire 90
4.5.2 Calcul le quantificateur flou par le degré de vérité 92
4.5.3 L'approche proposé IFlex- RLR-CardF par le degré de validité 93
4.5.4 Aperçue sur l'application 97
4.6 Approche proposé d'un classifieur médical « Classifieur- RLR-CardF»: 98
4.6.1 Schéma général de l'approche proposé « Classifieur- RLR-CardF » 98
4.6.2 Processus de développement, Classifieur- RLR-CardF 99
4.6.3 Aperçue sur l'application 104

4.7 Résultats et discussions 105

4.8 Comparaison de l'approche proposée avec les travaux de la littérature 111

4.9 Conclusion 112

Conclusion général*.....*113

Bibliographie*.....*115

Table des Figures

Figure 1-1 Schéma de la réduction sémantique de données 8
Figure 1-2Partition l'attribut Age 11
Figure 1-3Exemple d'arbre de hiérarchie [16] 16
Figure 3-1Les étapes de système de classification 51
Figure 3-2Hyperplan séparateur 53
Figure 3-3Arbre de décision 55
Figure 3-4Forme général de réseau de neurone 56
Figure 3-5Multi couche de réseau neurone 56
Figure 4-1 schéma général de notre modèle 80
4-2 Fenêtre Principale de notre application médical 81
Figure 4-3Représentation d'un sous-ensemble fou Jeune 84
Figure 4-4A partition flou des attributs Npreg,glu,Bp,SKIN 85
Figure 4-5A partition flou des attributs insulin,BMI,PED,Age 85
Figure 4-6 Représentation de fichier resumé-pima.Txt (RLR-CardF de la BD -PIMA) 87
Figure 4-7 Aperçus sur les entrés des paramètres des variables linguistiques 89
Figure 4-8 Aperçus sur le résumé linguistique RLR-CardF de la base PIMA 89
Figure 4-9 Graphe de comparaisons entre deux types de recherches 92
Figure 4-10Résultat de la recherche d'une requête par les trois méthodes de recherches ... 97
Figure 4-11 Aperçus sur la réponse d'une requête flou 97
Figure 4-12 Schéma général de notre approche de la classification supervisé 98
Figure 4-13Fenêtre de choix de simulateur de classification 104
Figure 4-14 Fenêtre de l'exécution de la simulateur de la base de teste 105
Figure 4-15Comparaisons performance de notre classifieur sur des bases de données médicales 107
Figure 4-16Le taux de classification Classifieur-RLR-CardF de la base de données PIMA de différents seuils de validation 108
Figure 4-17 taux de classification Classifieur-RLR-CardF de la base de données WBCD de différents seuils de validation 109
Figure 4-18 taux de classification Classifieur-RLR-CardF de la base de données Mamography de différents seuils de validation 109
Figure 4-19 Comparaisons le taux de classification de notre modèle Classifieur-RLR-CardF effectué sur des bases de données médicales 110
Figure 4-20 Histogramme de taux de classification de notre modèle Classifieur-RLR-CardF sur les bases de données 110

Table des tableaux

Tableau 1-1Réécriture de n-uplet(Fran et al., 2007) 14
Tableau 1-2Exemple de regroupement de données(Fran et al., 2007) 16
Tableau 1-3Exemple d'une base relationnelle (Pei et al., 2009) 18
Tableau 4-1 Caractéristiques des jeux de données utilisés 77
Tableau 4-2 Table de fuzzification de la base de donées PIMA 86
Tableau 4-3Tableau de α -cut de C_{α}^{1} and C_{α}^{2} 95
Tableau 4-4 Tableau de la fonction Graduel 95
Tableau 4-5 Base de connaissance 96
Tableau 4-6Fuzzification table of patient attributes 102
Tableau 4-7Table T1, T2 103
Tableau 4-8 Tableau de la similarité 103
Tableau 4-9Tableau de performance de Classifieur-RLR-CardF 106
Tableau 4-10 Les résultats obtenus par notre approche de différent seuil de validation. 108
Tableau 4-11Table de comparaisons notre approche de classification avec d'autres travaux de la litérattures 111

Introduction Générale

Contexte Général

Durant ces dernières années, la collecte des données médicales a donné lieu à une grande masse de l'information. Les bases de données médicales contiennent les informations sur les maladies et les pathologies des patients. Ces informations avec les techniques de l'intelligence artificiel ont permet à plusieurs chercheurs à développer des systèmes d'aide au diagnostic médical.

A cause de la fatigue et la pression durant le travail quotidien, le médecin peut mettre des erreurs de diagnostic. Pour prendre une bonne décision, il faut que le médecin utilise une grande quantité de connaissances comme : les définitions sur la maladie, les paramètres de diagnostic, les combinaisons de médicaments, les corrélations entre les maladies…. etc. D'où, l'intérêt d'un système de diagnostic médical pour d'aider le médecin de prendre une décision fiable et précise.

Le développement des systèmes médicaux à base de connaissances a permis une collaboration entre le médecin, l'ingénieur en biomédical et le patient. Mais, l'utilisation d'une grande masse de données médicales dans ces systèmes a causé plusieurs problèmes dans leur structure, leur stockage et leur analyse.

L'objectif principal de cette thèse est de proposer un modèle de diagnostic médical qui traite une grande masse de données médicales d'une manière automatique et interprétable en langage naturel. Pour cela Nous faisons appel aux techniques de l'Intelligence Artificielle (IA).

L'intelligence artificielle est un domaine en informatique qui tente de doter les machines de l'intelligence. Elle a été appliqué dans plusieurs domaines comme l'éducation, le commerce, … etc. L'extraction des connaissances d'une grande masse de données est l'étape principale de l'intelligence artificielle. Il faut que le résultat obtenu contienne toutes les informations essentielles pour une meilleure prise de décision.

Problématique

Il existe de nombreuses méthodes dites intelligentes appliquées à des systèmes de diagnostic médical.

Ces systèmes sont d'un intérêt majeur pour les experts de la santé mais ils souffrent de plusieurs problèmes comme :

- La taille volumineuse de bases de données médicales (Big Data) où ces données sont difficiles à gérer, à stocker, à analyser.

- L’absence de l’interprétabilité, en langage naturel, des données de la base médicale où le médecin trouve une difficulté d'interpréter les données numériques de la base de données médicale.
- La difficulté de l’utilisateur à effectuer une recherche d'informations pertinentes en langage naturel.
- l’absence de l’interprétabilité dans les résultats de système d’aide au diagnostic médical.
- la difficulté de créer un classifieur performant et moins couteux sur une base de données volumineuse.

Les données médicales sont stockées progressivement dans une base de données très large d’où la recherche de l’information et la prise de décision deviennent très difficile. Il faut trouver un moyen pour réduire le volume de données médicales en se basant sur une approche sémantique.

Le but principal de notre recherche est de développer un modèle médical intelligent, basé sur la théorie de l’ensemble floue. Au début pour réduire le volume de données, nous utilisons l’une des méthodes de réduction sémantique sur la base de données numérique. Puis, nous proposons un classifieur qui utilise ces résumés sémantiques afin d’établir un diagnostic fiable et interprétable.

Contributions

En général, la grande majorité des informations stockées dans les bases de données médicales ne sont pas utilisées dans la prise de décision d’où la recherche des données pertinentes est difficile.

Il faut chercher une méthode de la réduction des données qui combine entre le système de recherche sémantique des données et le système de classification supervisé médicale.

Il existe de nombreuses techniques permettant de réduire la masse de données. Nous citons les méthodes les plus répondues :

- compression physique.
- compression logique.
- réduction sémantique.

En faite, la méthode traditionnelle de réduction de données, implique une perte de la sémantique de l’information où le langage de ces résumés est incompréhensible pour l’utilisateur. Cependant, la création d’un résumé intelligent est liée à une formulation compacte et intelligible.

Dans notre thèse, nous utilisons l’une des méthodes de résumés linguistiques pour traiter le volume des données médicales et de réaliser un modèle de décision et de recherche médical qui puisse identifier la maladie d’une manière simple et transparente et d’établir une recherche sémantique des données médicales.

Le résumé linguistique est basé sur la théorie des ensembles flous, proposé par Zadeh [1].Elle offre une interprétation sémantique en langage naturel. La théorie des ensembles flous est utilisée pour définir des résumés linguistiques car elle permet d'obtenir une description linguistique des données. L'idée de base est de présenté les données par des termes linguistiques (comme presque tous, autour de nous) et d'effectuer des calculs sur une requête floue afin d'estimer son degré de vérité.

Dans la littérature, il existe de nombreux modèles de résumé linguistiques, nous citons en particuliers :

- le modèle SAINTETIQ[2]
- le modèle de Yager [3].
- le résumé à base de calcul de la cardinalité floue [4].

Dans le cadre de notre travail, nous avons choisi le résumé linguistique à base de calcul de cardinalité d'un ensemble flou par ce qu'il condense toutes les informations, avec leur poids et leurs degrés de satisfaction. Il enregistre les informations par des variables linguistiques pour qu'il soit proche du langage humain. Le cardinalité flou et le degré d'appartenance sont calculés pour chaque variable, pour rendre le résumé plus explicite et plus précis.

De ce faite, nous intéressons à la conception d'un modèle de diagnostic médical complet qui combine entre trois approches qui sont proposés dans notre travail de recherche :

RLR-CardF (**R**ésumé **L**inguistique **R**éduit à base de **c**alcul de la **Card**inalité **F**loue).

IFlex- RL (**I**nterrogation **Flex**ible des **R**ésumés **L**inguistiques).

Classifieur- RLR-CardF (**Classifieur** à base **RLR-CardF**).

Ainsi, ces trois approches, proposées dans notre thèse, sont définies ci-dessus :

RLR-CardF : Nous proposons de à utilisé un nouveau algorithme de la méthode de résumé linguistique, à base de calcul de la cardinalité floue, afin de faire une réduction sémantique de la base de données médicale.

IFlex- RL : nous proposons un algorithme interrogation flexible qui est basé sur le calcul de degré de validité , proposé par l.Liétard[5] , afin d'interroger nos résumés linguistiques médicaux et chercher la réponse exacte d'une requête, selon les préférences de l'expert du domaine. Nous traitons également deux types de requêtes, une requête floue simple (Q R sont T) et une requête floue complexe (**Q R et P sont T**) où **R** et **P** sont deux prédicats flous, **Q** est un quantificateur flou, **T** est une degré de vérité de cette requête . Dans notre système d'interrogation flexible, deux facteurs sont pris en compte, le facteur temps et le facteur de précision.

Classifieur- RLR-CardF: A l'aide de calculs de similarité, entre les résumés linguistiques, nous proposons d'adapter ces calculs pour créer un modèle de diagnostic médical, c. à. d,

proposons une nouvelle méthode classification supervisé. Nous obtenons des résumés linguistiques sous forme de base de connaissances (KB), que nous utiliserons dans notre Classifieur médical.

En définitive, notre modèle médical présente des résultats interprétables en langage naturel (logique flou). Il combine entre le système d'interrogation flexible et le classifieur supervisé.

Organisation du manuscrit

Afin de fournir au lecteur une vue d'ensemble du contenu de notre thèse, nous décrivons brièvement le contenu de chacun de ces chapitres:

Chapitre1 : Nous présentons les différentes techniques de la réduction de données .Ensuite, nous détaillons les différents modèles de résumé linguistique.

Chapitre2 : Nous définissons le système d'interrogation flexible. Puis, nous citons les différentes approches pour créer un système d'interrogation flexible pour chaque modèle de résumé linguistique.

Chapitre3 : Nous définissons, tout d'abord, les systèmes de classification supervisé .En suite, nous présentons un état de l'art sur les méthodes souvent utilisées dans les classifieur supervisés. Enfin, nous clôturons ce chapitre par un état de l'art sur les systèmes de déduction en utilisant les résumés linguistiques.

Chapitre4 : Tout d'abord, nous présentons l'implémentation de notre résumé linguistique **RLR-CardF** , sur plusieurs bases de données médicales. En suite, Nous créons un système d'interrogation médical, **IFlex-RL**, qui interroge nos résumés linguistiques médicaux. En outre, nous implémentons un système de classification médical .Nous proposons un nouveau type de classifieur, **Classifieur-RLR-CardF** , qui est basé sur le calcul de la similarité des résumé linguistiques. Enfin, nous nous consacrons à l'interprétation et les discussions des résultats, tout en décrivant les implémentations faites avec les résultats expérimentaux obtenus.

chapitre 1. Résumé linguistique

1.1 Introduction 6

1.2 Compression de données 6

1.2.1 Compression physique 6

1.2.2 Compression logique 7

1.3 Réduction sémantique de données 7

1.3.1 La réduction basée sur les méthodes statistiques 8

1.3.2 Approche basé sur les modèles 9

1.4 Résumé linguistique : 10

1.4.1 Définition Variable linguistique : 10

1.4.2 Théorie de l'ensemble flou 11

1.5 Résumé SaintEtiQ 13

1.5.1 Les étapes de la construction hiérarchique de résumés. 14

1.6 Résumé de Yager 16

1.7 Résumé à base de calcul de la cardinalité floue 20

1.7.1 Principe de résumé linguistique à base de calcul de la cardinalité floue 21

1.7.2 Définition le Quantificateur flou 24

1.8 Conclusion 25

1.1 Introduction

Les volumes de données stockés par les systèmes informatiques deviennent importants à cause de l'archivage de données numériques et de l'avancement technologique .Pour cela, la recherche des informations pertinentes, pour un besoin spécifique, devient difficile. Il faut trouver une technique de résumé qui puisse couvrir toutes les données, minimiser le volume des résumés et de les présenter sous une forme simple et intelligible.

Dans l'état de l'art, ils existent de nombreuses techniques de réduction de volume de données qui prennent en compte le traitement informatique. Parmi ces techniques, nous citons un modèle de résumé linguistique qui est très répondu dans la littérature où nous détaillons ses principes dans ce chapitre. Ce type de modèle de résumé est une méthode de construction d'un résumé de données structurées. Il est basé sur le concept de théorie de l'ensemble floue.

Mais avant de parler sur ces différentes méthodes de résumé linguistique, nous citons dans la section 1.2, section 1.3, un état de l'art sur les techniques de condensation de l'information et nous montrons la différence entre la compression de données et la réduction sémantique de données.

1.2 Compression de données

La compression de données est une action, utilisée pour réduire la taille physique d'un ensemble des informations dans le même espace de stockage, afin de les transférés. Nous pouvons dire que la compression permet de transférer les données de tous les formes de fichiers graphiques (les images, vidéo…) plus rapidement. Il existe pour cela deux types de compression de données : la compression physique, la compression logique. La différence entre les deux est de voir comment les données sont compressées puis réarrangées dans une forme plus compacte.

1.2.1 Compression physique

La compression physique est un algorithme capable de condenser les données dans un minimum de place.

Dans ce type de compression, les chercheurs ont utilisé plusieurs techniques de compression où la meilleure technique de compression destinée à la meilleure utilisation de l'espace disponible.

En est fait, le résultat d'un bloc de données compressées est plus petit que l'original car l'algorithme de compression physique a retiré la redondance qui existait entre les données elles-mêmes.

L'application la plus représentative de la compression de données c'est la compression de fichiers, utilisant des algorithmes (RLE, LZW, codage de Huffman etc.)[6] ou utilitaires (PkZip, 7-zip, deflate, compress, etc.) [7]

Avec la croissance de technologie en informatique, il faut plus d'espace de stockage de données et de bonne méthode de rangement .Dans les méthodes de compression physique de données, il n'y a plus besoin de trier pour faire de la place pour passer à une capacité supérieure. Mais cela cause un problème de traitements informatiques parce qu'elle implique une augmentation du volume de données à traiter.[8]

Aussi, cette méthode produit des résultats incompréhensibles qui apparemment n'ont aucun sens.

1.2.2 Compression logique

La compression logique est un algorithme qui permet de recoder les données dans une représentation différente et plus compacte, contenant la même information. Elle est accomplie à travers le processus de substitution logique qui consiste à remplacer un symbole alphabétique, numérique ou binaire en un autre. Par exemple : Changer "United State of America" en "USA où "USA" est dérivé directement de l'information contenue dans la chaîne "United State of America" donc nous gardons la même signification. Mais, La substitution logique ne fonctionne pas dans les informations de contenue exemple remplacer 1987 par 87....[9]

Les différentes méthodes de compression logique sont basées sur 3 critères :

– Le taux de compression : c'est le rapport de la taille du fichier compressé sur la taille du fichier initial.

– La qualité de compression : sans ou avec pertes (avec le pourcentage de perte).

– La vitesse de compression et de décompression.

Donc, la compression des données sert à réduire le volume des données où le meilleur algorithme de compression est lié à la capacité de stockage et au débit de transmission. Aussi, nous pouvons juger l'algorithme par la qualité de compression.

Dans ces types de compression, l'algorithme est dit « sans perte » parce que l'algorithme inverse, appliqué au résultat, retourne exactement les données initiales. Dans la compression « avec perte », pour certains types de données (audio, vidéo et photo), l'algorithme inverse de données ne garantit pas de trouver les mêmes données initiales.

Avec la croissance du volume des données à traiter, les chercheurs se sont dirigés vers d'autres voies de recherche, afin de réduire le volume de données sémantiquement.

1.3 Réduction sémantique de données

La réduction sémantique de données est de créer un résumé des données où la forme de résumé obtenu est différente de celle des données initiales. Donc, il faut chercher un moyen pour gérer le problème du volume des données en prenant en compte le traitement informatique.

Il existe plusieurs techniques de réduction de données sémantique qui effectuent des traitements sur l'ensemble des données, afin de construire un résumé significatif.

Nous pouvons résumer les différentes méthodes de la réduction sémantique de données par ce graphe ci-dessous (figure1.1). Ces différentes méthodes ont un objectif commun : c'est la réduction d'une grande masse de données structurées en respectant leur sémantique.

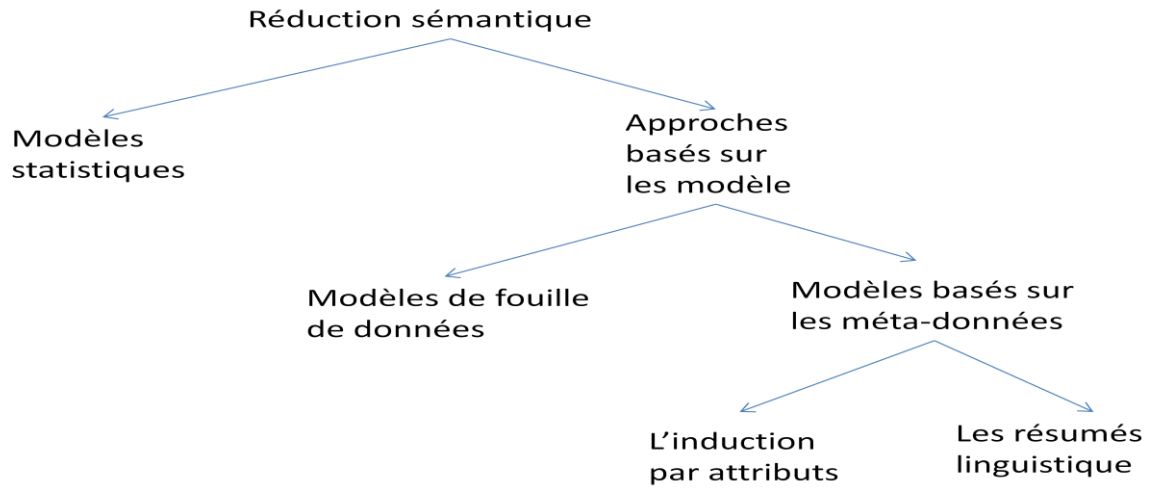


Figure 1-1 Schéma de la réduction sémantique de données

1.3.1 La réduction basée sur les méthodes statistiques

La méthode statistique est une partie du processus d'analyse de données qui construit un résumé à partir d'une grande masse de données. Son principe est d'extraire des informations les plus intéressantes. Alors, il faut trouver des algorithmes qui traitent les bases de données en prenant en compte la rapidité du temps de réponse et la quantité de données.

Les BDS(Bases de Données Statistiques), apparues au début des années 1970, ont comme principales caractéristiques, de collecter des informations et de les fournir aux utilisateurs des outils pour analyser des données. Dès les années 80, les travaux sur les BDS [10]sont consacrés à l'étude de la construction d'information, de la maintenance et de la pertinence d'agrégats, calculés à partir de fonctions statistiques. Les chercheurs ont choisi de représenter les données et de les traiter d'une manière résumée (par groupe d'individus ou par sous-ensembles). Les données dans les BDS sont de deux types de valeurs : ceux qui représentent les paramètres afin d'identifier des catégories de données, et ceux qui représentent les variables qui sont appelées les mesures.

Il existe trois types de méthodes statistiques utilisées dans l'analyse de données :

– La méthode verticale basée sur la réduction du nombre d'instances.

– La méthode horizontale basée sur la réduction du nombre d'attributs.

– La méthode du calcul d'agrégat.

1.3.1.1 La réduction verticale

Nous réduisons le nombre d'instances qui sont choisies parmi les données initiales. Nous utilisons pour cela l'échantillonnage, les histogrammes ou les techniques de classification [11][12].Leur principe est de transformer les données dans un autre mode de représentation .

Définition de l'échantillon *: Un échantillon est un sous-ensemble d'une population sur laquelle, on effectue une étude statistique*[8]

1.3.1.2 Réduction horizontale

Cette approche est basée sur le résumé du nombre d'attributs. Elle consiste à réduire la complexité des données en ne gardant qu'un nombre restreint d'attributs et les plus importants. La méthode d'analyse factorielle est la méthode la plus utilisée. Car les analyses factorielles trouvent tout leur intérêt pour la compréhension des tableaux de grande dimension.par contre, les traitements statistiques classiques ne peuvent interpréter les grandes tableaux de données .Cette méthode a été utilisée pour la réduction du nombre d'attributs d'une base de données[8].

Pour cela , nous utilisons les méthodes de corrélations entre les attributs pour éliminer certain attributs et garder les plus importants[13][10] . Les plus courantes des méthodes factorielles sont :

- Analyse en Composantes Principales (ACP),
- Analyse Factorielle des Correspondances (AFC)
- L'analyse discriminante.

1.3.1.3 Calcul d'agrégats

Les approches de compression sémantique, s'appuyant sur le calcul d'agrégats se trouvent dans les BDS (Bases de Données Statistiques) et dans les systèmes OLAP(Online Analytical Processing). Les opérations d'agrégation et de désagrégation sont définies pour représenter les informations à différents niveaux de granularité. Ces agrégations réduisent aussi l'espace de stockage des données, comme le fait de ne stocker que des codes à la place des valeurs des catégories. Le calcul d'agrégats dans les BDS a permis la création des tables résumées décrivant la distribution multi variée des données [14][8] .

1.3.2 Approche basé sur les modèles

Il existe deux types d'approches basé sur les modèles.

1.3.2.1 Le modèle de fouille de données

La fouille de données consiste à chercher et à extraire l'information (utile ou inconnue) dans les gros volumes de données, qui sont stockées dans des bases ou des entrepôts de données. Elle est liée avec l'apprentissage automatique mais la difficulté se trouve au niveau des données initiales. Ces données sont des données complexe, hystérogènes et

volumineuses. Nous citons quelque d'applications qui utilisent la notion de fouille de données (Amazon, lastfm, netflix…ect)

1.3.2.2 Les modèles basés sur les métadonnées

Ces modèles sont considérées comme des approches de fouille de données descriptive car ils sont basés sur la description des concepts où ils décrivent les données (les concepts) pertinentes pour analyser sous une forme réduit et compréhensible, afin d'obtenir une forme concise et générale.

Il existe deux grands modèles de la réduction basés sur les métadonnées : l'induction par attribut (par exemple l'arbre de décision) et les résumés linguistiques. Nous nous intéressons uniquement dans notre thèse aux modèles basés sur les résumés linguistiques où nous allons détailler dans la section 1.4 ce type de résumé sémantique.

Enfin, Il existe d'autres modèle pour la réduction sémantique de données ,comme le système caches digest (distribués en mode pair à pair).Il est considérer comme un modèle de la réduction de données et il se retrouve dans les caches Internet [15].C'est une synthèse du contenu d'un cache Web, par exemple les moteurs de recherche est les utilisée pour répondre aux requêtes (Google , Yahoo Search,…) [7].

1.4 Résumé linguistique :

Les résumés linguistiques se basent sur l'exploitation des variables linguistiques introduites par L. Zadeh en 1965(Zadeh, 1965) afin de construire un résumé compréhensible en langage naturel et complet. Ces méthodes de résumés linguistiques existantes utilisent la théorie des sous-ensembles flous. Elles décrivent les données par des termes « linguistiques » pour, à la fin, les résumer produits sont appelés les « résumés linguistiques ».

1.4.1 Définition Variable linguistique :

Une variable linguistique est représentée par un (V, Ω, L_V)

Où V est une variable (par exemple salaire, l'âge, ...) définit sur un ensemble de référence $\Omega(\mathrm{R}, \mathrm{N}, ...)$ et que la valeur peut être un élément de Ω.

On note L_V = {X; Y; ...} un ensemble de sous-ensembles flous de Ω, nous utilisons L_V pour caractériser le variable V .

Un élément de L_V est aussi appelé étiquette linguistique.

Afin de faire une partition floue, les variables linguistiques permettent de partitionner le domaine de l'attribut en étiquettes linguistiques qui sont associées avec un degré d'appartenance. Par exemple la partition de l'attribut l'âge est réalisée par deux partitions floues qui sont décrites à la figure 1.2.

Figure 1-2Partition l'attribut Age

Exemple : Nous avons l'enregistrement (Ali, 45) alors la fuzzification de cette enregistrement selon le figure1.2 est :

Selon le graphe, V représente Age des individus .Nous avons deux variables linguistiques (Agé, jeune).

Donc, nous présentons par :

(Ali ,0.3/$_{\text{jeune}}$ et 0.7/$_{\text{Agé}}$) C.à.d. qu'Ali appartient partiellement au variable linguistique jeune (Agé) avec un degré d'appartenance 0.3 (0.7).

Les partions flous se fait par la théorie de l'ensemble flou.

1.4.2 Théorie de l'ensemble flou

La théorie de sous-ensemble flou, proposée par Zadeh [1]est un outil mathématique qui définit le concept d'appartenance partiel d' un ensemble et offre une solution aux problèmes de modélisation de donnée.

Grace à cette théorie, nous exprimons plus facilement les transitions graduelles, les informations imprécises ou les classes mal définies. Il devient possible d'exprimer les données numérique en langage naturel où nous utilisons les termes linguistiques, par exemple : grand, jeune, élevé, …etc.

En effet, les ensembles flous ont été présentés pour la première fois en 1965 par L. Zadeh. Ils se présentent comme une extension des ensembles usuels, ils ont des caractéristiques spécifiques de domaine de valeur.

Dans cette théorie, un ensemble produit au lieu d'être une valeur d'un l'ensemble de deux éléments binaires {0,1}, il est défini sur l'intervalle [0,1].

Un sous ensemble A de X est défini par une fonction d'appartenance qui est associé à chaque élément x de X, le degré $f_A(x)$, compris entre 0 et 1 avec lequel x appartient à A [1]:

$$f_A : X \rightarrow [0,1] \qquad (1.1)$$

Où $f_A(x)$ est une fonction d'appartenance (caractéristique) d'un ensemble flou A .

Nous trouvons également dans cette théorie fondamentale, l'extension d'ensembles flous du couple intersection/union, appelé aussi : norme et co-norme triangulaire.

La norme triangulaire : est une norme mathématique définie sur l'intervalle [0,1] et qui satisfait les axiomes de commutativité, associativité, monotonie et possède comme élément neutre 1, elle est notée par " $\wedge$ ".

La co-norme est aussi une norme vérifiant les mêmes axiomes, mis à part l'élément neutre qui est 0 et non pas 1, elle est notée par :" $\vee$ ".

Ces deux opérateurs sont liés entre eux par la loi :

$$(a \wedge b) = 1 - (1-a \vee 1-b) \qquad (1.2)$$

Où a,b sont deux ensemble flou.

Même, Il existe des normes et co-normes particulières :

$$(a \wedge b) = min(a,b) \text{ et } (a \vee b) = max(a,b) \qquad (1.3)$$

Ce couple présente une particularité puisque la norme est la plus grande des normes et la co-norme est la plus petite des co-normes.

Aussi, nous pouvons présenter le couple par :

$$(a \wedge b) = a*b \text{ et. } (a \vee b) = a + b - a*b \qquad (1.4)$$

Il existe aussi d'autres opérations sur les ensembles flous comme la négation ou le complémentaire.

Donc, une partition floue correspond à l'extension floue de la partition de l'espace dans les ensembles classiques. Elle est une division d'un ensemble en sous-ensembles flous. Avec l'ajout d'une contrainte à cette partition, nous obtenons des informations plus exploitables. Elle est présente par:

$$\forall x \in E \sum_i \mu_i (x) = 1 \qquad (1.5)$$

Où μ_i est une fonction d'appartenance pour le sous-ensemble i. Cela signifie qu'un élément de l'ensemble ne peut appartenir qu'à deux sous-ensembles au maximum.[16].

Donc, le point commun des méthodes de résumé linguistique est d'exploiter la théorie de sous-ensemble flou, proposée par Zadeh , ceci pour exprimer des concepts vagues et décrire les valeurs d'attributs par des termes linguistiques.

Par exemple : « les personnes jeunes et glucose élevé ». Les enregistrements appartiennent à un ou plusieurs concepts ,suivant l'adéquation entre ses valeurs d'attribut et les termes linguistiques (jeune et élevé).On parle ici sur la présentation de résumé linguistique de Yager [17].

Prenons un exemple simple, dans une base de données relationnel qui contient plusieurs attributs :

- un ensemble des entiers pour l'attribut Age.
- La technique de résumé consiste à modifier le domaine initial en un nouveau domaine plus imprécis, par exemple {jeune, âgé}.
- La base de données est réécrite avec ce nouveau « vocabulaire » .où les t_uplets devenus terme linguistique, associés à des valeurs de l'apprentissage et de la vérité.

Alors , nous déduisons que le choix des nouveaux domaines d'attributs est une étape clé du processus de résume. Nous pouvons classer le résumé linguistique dans trois catégories : le résumé quantifié (résumé de Yager), le résumé à base de calcul de cardinalité floue et le résumé structuré en une hiérarchie (résumé SaintEtiq).

Dans notre thèse, nous nous intéressons au résumé à base de calcul de cardinalité flou, proposé par Dubois et Prade[18]. Cette méthode utilise la notion de cardinalité pour chaque sous ensemble floue où elle permet de combiné entre la notion de qualité et la notion de quantifier de la requête, ce qui a était absent dans les autres méthodes de résumé linguistique.

Dans les sections suivantes, nous présentons les trois types de résumé linguistiques.

1.5 Résumé SaintEtiQ

L'objectif de modèle SaintEtiQ[19] est de donner une vue synthétique d'une base de données. Ceci est réalisé à travers de résumés linguistiques structurés en une hiérarchie. Dans ce modèle, l'utilisateur définit les connaissances de domaine, il fourni un vocabulaire pour résumer et le rendre intelligible.

Dans ce type de résumé, il y'a deux types d'informations d'entrées :

- **Les données à résumer** : elles sont des données relationnelles qui se présentent par des t_uples , elles suivent le schéma d'une relation R définie sur un ensemble d'attributs A = { $A_1, A_2, \ldots, A_n$ }.
- **Les connaissances de domaine** : elles donnent des indications sur la manière de résumer les données.

Chaque attribut A_i a un domaine de définition D_{A_i} et un enregistrement t qui est un t_uplet formé de valeurs suivant l'ordre prédéfini des attributs Ai. L'ensemble des t_uplet est noté par:

$(t.A_1, t.A_2, \ldots, t.A_n)$.[7]

1.5.1 Les étapes de la construction hiérarchique de résumés.

1.5.1.1 Réécriture des données :

Cette étape a pour objectif de réécrire une base de données numérique pour obtenir une vue simplifiée. Cela consiste à changer l'écriture de données à une forme compréhensible et vague C.à.d. formuler les valeurs d'attributs (numériques) à des variables linguistiques, qui sont associé avec un degré d'appartenance.

. Par exemple, pour les n-uplets d'une base, la valeur d'âge d'une personne sera remplacée par une valeur linguistique, comme jeune.

Exemple : Considérons un enregistrement (Ali, 39) . Cherchons la variable linguistique appartenant à cette t_uplet .La valeur d'attribut Age peut prendre la variable linguistique **jeune** avec un degré de $0,8$ $(\mu_{jeune}(28) = 0,8)$ et la variable linguistique **âgé** avec un degré de $0,2$ $(\mu_{agé}(28) = 0,2)$. Ceci s'écrire de la manière suivante :

$$t : (\text{Ali}, 28) \rightarrow \begin{matrix} t1 : (\text{Ali}, 0{,}8/\text{jeune}) \\ t2 : (\text{Ali}, 0{,}2/\text{agé}) \end{matrix}$$

Ensuite, le n-uplets se réécrit en plusieurs exemplaires. Autant de fois qu'il y a une association possible provenant de la reformulation des valeurs d'attributs. Donc, avec 2 attributs réécrits, il est possible d'obtenir 4 n-uplets réécrit.

Nous présentons un exemple cité dans[16], il explique ,comment construire le résumé par la méthode SaintEtiq.

Le tableau 1.1 proposé par [16] montre les résultats obtenus après la réécriture des n-uplets (UZ40, 1, 38,900), (CuSn12, 8, 40, 850), (CuAsO5, 12,44, 896), (Fe, 10, 35, 1530) et (Ni, 5, 35, 1453) suivant les partitions définies dan[20]. Nous constatons que CuSn12 et CuAsO5 ont été réécrit chacun deux fois.

Tableau 1-1Réécriture de n-uplet[16]

Matériaux	Epaisseur		Dureté		Température		Réf
Ta1	Moyenne	0.70	Tendre	1.00	Modérée	0.85	UZ40
Tb1	Moyenne	0.35	Tendre	0.90	Modérée	1.00	CuSn12
Tb2	Mince	0.35	Tendre	0.90	Modérée	1.00	CuSn12
Tc1	Moyenne	1.00	Tendre	0.40	Modérée	0.90	CuAs05
Tc2	Moyenne	1.00	Dure	0.40	Modérée	0.90	CuAs05
Td1	Moyenne	0.70	Tendre	1.00	Normale	0.85	Fe
Te1	Mince	1.00	Tendre	1.00	Normale	0.96	Ni

D'une manière générale, nous considérons que le résultat de la réécriture est le premier niveau de la hiérarchie de résumés. A l'aide des connaissances de domaine, cette phase a pu l'interpréter des valeurs d'attributs à des variables linguistiques. Ces connaissances sont constituées essentiellement par des variables linguistiques, définies sur les domaines d'attributs de la relation résumée. Ce sont les utilisateurs ou les experts qui proposent ces connaissances de domaine. Leurs avantages est de définir un langage de description de données sémantique qui est le plus proche possible de l'humain. Les connaissances de domaine permettent aussi de faire la correspondance entre les valeurs de domaines d'attribut et le vocabulaire d'expression des résumés de données. Un domaine D_A est alors réécrit par l'ensemble, noté $D + A$.

1.5.1.2 Regroupement des données

L'objectif du regroupement de données est de les organiser afin d'obtenir une représentation sous forme d'arbre. Chaque niveau de cet arbre présente un niveau de résumé différent.

Pour regrouper les données, plusieurs stratégies peuvent être envisagées:

- Il faut demander à l'utilisateur de fournir des étiquettes linguistiques de niveau supérieur afin de regrouper les données qui sont proches des besoins de l'utilisateur.
- Nous pouvons regrouper les données en fonction de leur similarité sans demander les informations à l'utilisateur.
- Pour chaque n-uplet candidat qui doit s'ajouter dans une hiérarchie (arbre) grandissante. Ensuite, il faut chercher une feuille d'un autre candidat qui a la même étiquette linguistique.

Lors de la construction de l'arbre, la structure de l'arbre évolue continuellement. Cette évolution est contrôlée partiellement par les opérateurs d'apprentissage, présentés dans [20], qui cherchent à maximiser les similitudes à l'intérieur d'une classe et les dissemblances entre les classes.

Le tableau 1.2 représente le regroupement qui peut être obtenu à partir des données qui ont été réécrites précédemment (tableau1. 1)[20]. On peut remarquer que dès le premier niveau ,certains n- uplets issus de la réécriture sont regroupés comme suit :

Tableau 1-2Exemple de regroupement de données[16]

Résumé	Intention	Recouvre
Z0	{1.0/mince+1.0/Moyenne+0.7/Large…..}	Z1,Z2
Z1	{1.0/moyenne+1.0/tendre….}	Z3,Z4
Z2	{1.0/mince+1.0/Moyenne +…}	Z5,Z6,Z7
Z3	{1.0/moyenne+1.0/tendre….}	Ta1,tb1,tc1
Z4	{0.75/moyenne+1.0/tendre….}	Td1
Z5	{0.35/mince+0.9/tendre+…}	Tb2
Z6	{1.0/moyenne+0.4/dure….}	Tc2
Z7	{1.0/mince+1.0/tendre+…}	Te1

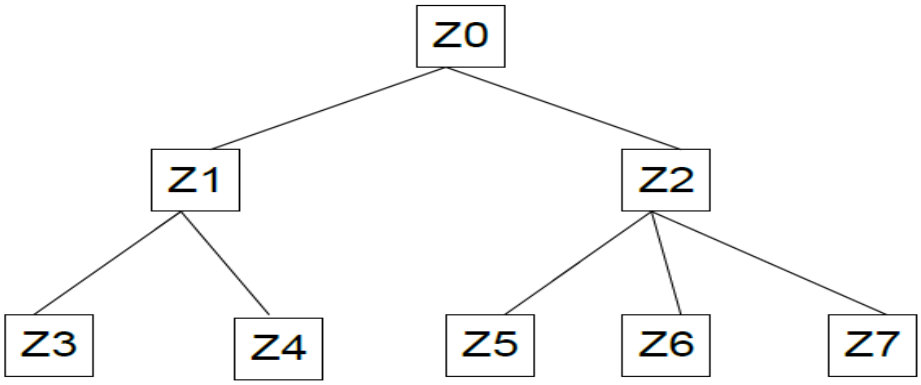


Figure 1-3Exemple d'arbre de hiérarchie [16]

Donc, nous constatons que chaque niveau de la hiérarchie permet d'avoir une vue d'ensemble de la base d'origine, ainsi que les niveaux les plus faibles qui contiennent les résumés les plus spécifiques.[16].

Malgré la vue hiérarchie de la méthode SaintEtiq mais la recherche des requêtes reste limiter. Cette méthode ne représente pas ces degrés d'appartenances dans l'arbre hiérarchie où il est impossible de rechercher des informations qualificatives.

1.6 Résumé de Yager

Le résumé linguistique quantifié proposée par Yager[17]utilise la notion de théorie de la Logique Floue sous une forme simple . Il utilise les variables linguistiques des ensembles de données. Dans l'état de l'art [21] et [22],ils présenté ce résumé par une forme plus avancée et applicable.

D'une façon générale, les résumés quantifiés [23] utilisent des quantificateurs flous pour décrire les données. Par exemple, dans le résumé de Yager [23], le résumé est présenté par :« la plupart des employés sont jeunes ». Le degré de satisfaction se fait par la mesure dans laquelle la proposition est satisfaite par les données.

Nous présentons ci-dessus le processus général de la création de résumé linguistique de Yager :

Nous avons:

$\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ est un ensemble enregistrements dans une base de données, par exemple : l'ensemble des patients.

$A = \{A_1, \ldots, A_m\}$ } est un ensemble d'attributs qui caractérisent les objets de Y, par exemple :le Glucose(GLU), l'âge, etc.

$A_j(y_i)$ représente une valeur d'attribut A_j pour objet y_i. [24]

Alors, un résumé linguistique d'un ensemble de données se compose de quatre éléments essentiels:

- **un résumer** S : c.à.d. un attribut avec une variable linguistique (prédicat floue) définie sur le domaine de l'attribut A_j (par exemple «**faible**» pour l'attribut «**GLU**»);
- **une quantité** Q : quantificateur linguistique (par exemple **la plupart**);
- **vérité** (validité) T du résumé : c'est un nombre qui appartient à l'intervalle [0, 1] et qui évalue la vérité (la validité) du résumé (par exemple 0,7); les résumés avec une valeur élevée de T sont intéressants;
- R , est un autre attribut avec une variable linguistique (prédicat floue) définie sur le domaine de la détermination d'un attribut A_k (sous-ensemble flou) de Y (par exemple «jeune» de l'attribut «âge»).

Les résumés linguistiques peuvent être illustrés par :

$$T\ (most\ of\ \text{patients}\ earn\ low\ \text{Glu}) = 0.7 \qquad (1.6)$$

Ou

$$T\ (most\ of\ young\ \text{patients}\ earn\ low\ \text{Glu}) = 0.7 \qquad (1.7)$$

Le prototype proposé dans ce résumé est le même qui a était proposé par Zadeh [25] correspondant à l'une ou l'autre des équations (1.6) et (1.7).

D'une façon générale, nous trouvons le résumé linguistique du type simple qui contient un seul prédicat flou :

$$Qy's\ are\ S \qquad (1.8)$$

Et le résumé complexe qui contient deux prédicats floues :

$$QRy's\ are\ S \tag{1.9}$$

Le calcul de valeur de vérité T se fait soit par le calcul proposé par Zadeh (quantificateur linguistique)[26], ou le calcul proposé par Yager [27]

Supposons que le quantificateur linguistique Q est un ensemble flou dans [0, 1] .Alors, les valeurs de T sont calculées par :

$$T(Qy's\,are\,S) = \mu_Q\left[\frac{1}{n}\sum_{i=1}^{n}\mu_S(y_i)\right] \tag{1.10}$$

$$T(QRy's\,are\,S) = \mu_Q\left[\frac{\sum_{i=1}^{n}\left(\mu_R(y_i)\wedge\mu_S(y_i)\right)}{\sum_{i=1}^{n}\mu_R(y_i)}\right] \tag{1.11}$$

Exemple :

Ce tableau représente une base des patients qui contient deux attributs (âge, Glu) :

Tableau 1-3Exemple d'une base relationnelle [24]

V\Y	**y1**	**y2**	**y3**	**y4**	**y5**	**y6**	**y7**	**y8**	**y9**	**y10**	**y11**
Age	25	48	31	35	28	51	37	43	34	27	53
Glu	1.8	2.0	2.8	3.0	2.8	3.0	2.3	2.5	3.5	2.9	3.0

Les experts du domaine ont fait les partions d'attributs pour obtenir des variables linguistiques : Sage={jeune(y), moyen-âge(ma)}, SGlu={faible(l),élever(h)}.Ils ont proposé trois quantificateurs flous : Q={plusieurs(s), environ_moitier(ah),la plupart(m)} .

Ainsi que :T={aproximativement_vrais(at),vrais(t),tous ta fait vrais(vt)}

Et, $\mu_y(x) = \begin{cases} 1 & if\ x \in [25,30] \\ 4 - \frac{1}{10}x & if\ x \in (30,40] \\ 0 & if\ x > 0 \end{cases}$ $\mu_{ma}(x) = \begin{cases} 1 & if\ x \geq 45 \\ \frac{1}{10}x - 3.5 & if\ x \in (35,45] \\ 0 & if\ x \leq 35 \end{cases}$

$\mu_l(x) = \begin{cases} 1 & if\ x \in [1.8,2] \\ \frac{1}{2}(2.5-x) & if\ x \in (2,2.5] \\ 0 & if\ x \geq 2.5 \end{cases}$ $\mu_h(\mathrm{x}) = \begin{cases} 1 & if\ x \in [3.3,3.5] \\ x - 2.3 & if\ x \in [2.3,3.3] \\ 0 & if\ x \prec 2.3 \end{cases}$

$$\mu_s(x)=\begin{cases}\frac{1}{2}\left(|x|-1\right) & if\ |x|\in[1,3]\\ 2-\frac{1}{3}|x| & if\ |x|\in(3,6]\\ 0 & if\ |x|>6\end{cases}\quad \mu_{ah}(x)=\begin{cases}0 & if\ |x|\prec 4\\ \frac{1}{2}|x|-2 & if\ |x|\in[4,6]\\ 4-\frac{1}{2}|x| & if\ |x|\in(6,8]\\ 0 & if\ |x|>8\end{cases}$$

$$\mu_m(\mathrm{x})=\begin{cases}1 & if\ |x|\in[10,12]\\ \frac{1}{4}\left(|x|-6\right) & if\ |x|\in[6,10)\\ 0 & if\ |x|\prec 6\end{cases}\quad \mu_{at}(x)=\begin{cases}1 & if\ x\in[0.5,1]\\ \frac{1}{2}x & if\ x\in[0,0.5)\end{cases}$$

$$\mu_t(x)=\begin{cases}1 & if\ x\in[0.8,1]\\ \frac{10}{3}(\mathrm{x}-0.5) & if\ x\in[0.5,0.8)\\ 0 & if\ x\in[0,0.5)\end{cases}\quad \mu_{vt}(x)=\begin{cases}5x-4 & if\ x\in[0.8,1]\\ 0 & if\ x\in[0,0.8)\end{cases}$$

Exemple, nous voulons vérifier la validité de résumé '' **Environ la moitié des patients ont un Glu élever ''.**

Donc :

1) Fixation s'=jeune s ''=élevée $\theta=0.5$

$$D_{s'}^{0.5}=\{V(y_i)|\mu_{s'}(V(y_i))\geq 0.5\}=\{25,31,35,28,34,27\}$$

$$D_{s''}^{0.5}=\{V(y_i)|\mu_{s''}(V(y_i))\geq 0.5\}=\{2.8,3.0,3.5,2.9,3.1\}$$

$$A_{s'}=\{y_i|V(y_i)\in D_{s'}^{0.5}\}=\{y_1,y_3,y_4,y_5,y_9,y_{10}\}$$

$$A_{s''}=\{y_i|V(y_i)\in D_{s''}^{0.5}\}=\{y_3,y_4,y_5,y_6,y_9,y_{10},y_{11},y_{12}\}$$

2) Sélectionner par μ_s,μ_{ah},μ_m , puis on obtient $\mu_s(A_{s'})=\mu_m(A_{s'})=\mu_{ah}(A_{s'})=1$

$$\max\{\mu_s(A_{s'}),\mu_m(A_{s'}),\mu_{ah}(A_{s'})\}=\mu_{ah}(A_{s'}) \qquad (1.12)$$

Selon μ_{at},μ_t,μ_{vt} et (1.12) :

$$\mu_{at}(\mu_{ah}(A_{s'}))=\mu_t(\mu_{ah}(A_{s'}))=\mu_{vt}(\mu_{ah}(A_{s'}))=1 \qquad .(1.13)$$

3) Sélectionner degré de vérité de $\mu_{vt}(\mu_{ah}(A_{s'}))$

Nous concluons que le résumé obtenu est valide :

Environ la moitié des patients ont un Glu élevée est tout à fait vrais.[24]

Malgré que le résumé de Yager est la méthode de résumé linguistique la plus utilisé mais L.Liétard[28] a montré ces limite. Car, le calcul du degré de vérité ne montre pas suffisamment la relation entre deux aspects(les informations quantitatives et les informations qualificatives). En plus, L'exécution cette méthode est couteuse par ce qu' il faut proposé tous types de requêtes puis vérifié les degrés de vérités de ces requêtes.

1.7 Résumé linguistique à base de calcul de la cardinalité floue

Comme les autres résumés linguistiques cités auparavant, le résumé à base de calcul de la cardinalité floue est basé aussi sur la théorie de l'ensemble flou où il transforme chaque enregistrement numérique à des variables linguistiques. Le principe de résumé est de compter le nombre des individus pour chaque variable linguistique.

Il existe plusieurs pistes de recherche pour représenter des cardinalités d'ensembles flous.

Au début , L.Zadeh[26], a présenté le calcul de sigma-count .Le principe repose sur le calcul de la somme des degrés d'appartenance de chaque élément de l'ensemble. Cette méthode, de représenter la cardinalité flou, pose des problèmes parce qu'on ne peut pas différencier les variables linguistiques qui ont un degré fort ou un degré faible.

Par exemple, la cardinalité, pour un élément , est présenté par le degré 1 .par contre pour dix éléments, la cardinalité est présenté dans un ensemble avec un degré 0,1.

Il existe une autre méthode proposé par par D. Rocacher dans[29] où ils ont pu régler le problème de Sigma Count. celle ci consiste à représenter la cardinalité d'un ensemble flou par un nombre graduel qui est basé sur la définition de la cardinalité floue ,définie par :

$$\forall n \in R, \mu_{card(E)}(n) = \sup\left\{\alpha \middle| \ddot{}card(E_\alpha) \geq n\right\} \text{ . .....(1.14)}$$

Ou :

- ✓ Le degré α est compris entre 0 et 1 (coupe)
- ✓ E un ensemble flou.

Leur principe est de représenter la cardinalité pour chaque degré d'appartenance de l'ensemble flou.

Par exemple, nous avons un ensemble flou A = {1/x1, 1/x2, 0.8/x3, 0.7/x4, 0.7/x5, 0.5/x6}, notons l'ensemble de cardinalité floue par card(A) = {1/0, 1/1, 1/2, 0.8/3, 0.7/4, 0.7/5, 0.5/6}.

Nous déduisons qu'on a 2 éléments qui appartiennent totalement à l'ensemble flou A et il y'a 1 élément qui appartient partiellement à l'ensemble A avec un degré de 0,8 .

Donc, cette représentation de la cardinalité permet de connaître pour chaque coupure de niveau (α- coupe) le nombre d'éléments de l'ensemble A qui ont un degré supérieur à α.

Définition : α-coupe ou coupe de niveau α est l'ensemble des éléments d'un ensemble flou où la valeur du degré d'appartenance est supérieure ou égale à α. (α est défini sur l'intervalle [0,1].)

En générale, le résumé à base de calcul de cardinalité floue est une méthode présentée dans [18]et qui se développe en deux phases.

La première phase, c'est l'étiquetage (fuzzification) où on remplace les enregistrements (n-uplet) par des variables linguistiques qui sont associé à un degré d'appartenance. Le but de cette phase est de constituer des éléments utiles pour la réalisation ou l'évaluation de différents types de résumés

La deuxième phase, nous commençons de calculer la cardinalité floue. Le but de cette dernière est de calculer le nombre de n-uplets d'une relation qui a la même étiquette floue. Et qu'on doit prendre en compte toutes les combinaisons d'étiquettes floues qui apparaissent dans la phase de réécriture des n-uplets.

1.7.1 Principe de résumé linguistique à base de calcul de la cardinalité floue

Soit r la relation des attributs A,B,C[4]

Soit (a_i, b_j, c_k) t_uplet de $r(A,B,C)$ projection d'attributs A,B et C.

Soit D_A, D_B, D_C domaines d'attribut. Nous supposons que $(A_1, A_2, ..., A_{na})$, $(B_1, B_2, ..., B_{nb})$, $(C_1, C_2, ..., C_{nc})$ sont les partitions floues de chaque domaines.

Nous supposons que la fonction d'appartenance des ensembles flous est sous forme de trapèze, aussi qu'il y a une échelle finie ($m+1$ niveau), on utilise, pour évaluer fonction d'appartenance,

$1 = \sigma_1 > ... > \sigma_m > 0$, où chaque niveau correspond à différente compréhension de A_r.

Dans la relation r, nous construisons une nouvelle relation r_{su} (pour "r résumé") par une procédure impliquant deux étapes principales :

- Etape 1 : Nous remplaçons chaque t_uplet $< a_i, b_j, c_k >$ par un ou plusieurs t_uplet de ensemble flou $< A_r, B_s, C_t >$ ou : $A_r(a_i) > 0, B_s(b_j) > 0, C_t(c_k) > 0$

Ainsi $< a_i, b_j, c_k >$ peut être remplacé par un t_uplet $< A_r, B_s, C_t >$.

Supposons qu'on a pour chaque attribut deux sous ensemble floue donc nous aurons 8 sous ensembles flous ($2^3 = 8$). Il faut que pour chaque t_uplet appartient au moins à un sous ensemble flou de même attribut.

- Etape2 : fusion et le calcul de cardinalité flou

Nous voulons savoir combien de t_uplet de r sont A_r, sont B_s, sont C_t, sont A_r et B_s, ..., sont A_r et B_s et C_t, pour toutes les étiquettes floues.

Pour avoir une représentation plus exacte de la relation r, nous utilisons des cardinalités floues au lieu de produit scalaires. Il est alors nécessaire de calculer les cardinalités de différentes étiquette linguistique et la conjonction entre ces étiquettes.

Tous les t_uplet de la forme $< x/A_r, y/B_s, z/C_t >$ qui sont identiques aux trois étiquètes, sont fusionnés en un seul t_uplet $< A_r, B_s, C_t >$ des r_{su}. En même temps, nous calculons les cardinalités de $F_{A_r}, F_{B_s}, F_{C_t}, F_{A_rB_s}, F_{A_rC_t}, F_{B_sC_t}, F_{A_rB_sC_t}$ où F_{A_r} (esp. $F_{B_s}, F_{C_t}, F_{A_rB_s}, F_{A_rC_t}, F_{B_sC_t}, F_{A_rB_sC_t}$) est un ensemble flou A_r (resp. B_s, C_t, A_r et B_s, A_r et C_t, B_s et C_t, A_r et B_s et C_t) défini sur les entiers {0,1,…} qui représente le nombre flou de t_uplet .

Expliquons le processus de la création des cardinalités floues :

Au début F_{A_r} = 1/0.

$$F_{A_r} = 1/0 + \ldots + 1/n - 1 + 1/n + \lambda_1/(n+1) + \ldots + \lambda_k/(n+k) + 0/(n+k+1) + \ldots \quad (1.16)$$

la valeur actuelle de la cardinalité floue F_{A_r} avec $1 > \lambda_1 \geq \ldots \geq \lambda_k > \lambda_{k+1} = 0$ et $n \geq 0$, $k \geq 0$.

Rappelons que cette expression représente une cardinalité qui prouvent au moins égale à n avec le degré 1 et au moins égale (n + k) avec le degré λ_k

Considérons un nouveau t_uplet dont la valeur A réécrit A_r. Soit x' est le degré d'appartenance de A_r dans ce t_uplet. Nous devrons alors modifié F_{A_r} .

Si x' = 1, le F_{A_r} devient:

$$F_{A_r} = 1/0 + \ldots + 1/n + 1/(n+1) + \lambda_1/(n+2) + \ldots + \lambda_k/(n+k+1) + 0/(n+k+2) + \ldots \quad (1.17)$$

Si x' <1, il y a deux cas : Soit $\exists i,\ x' = \lambda_i$ ou non.

Si $\exists i,\ x' = \lambda_i > \lambda_i + 1$ donc:

$$F_{A_r} = 1/0 + \ldots + 1/n - 1 + 1/n + \lambda_1/(n+1) + \ldots + \lambda_i/(n+i) + \lambda_i/(n+i+1) + \ldots + \lambda_k/(n+k+1) + 0/(n+k+2) + \ldots$$

(1.18)

si $\exists j, \lambda_j > x' > \lambda_j + 1$ donc: $\lambda_0 = 1$ et

$$F_{A_r} = 1/0 + \ldots + 1/n - 1 + 1/n + \lambda_1/(n+1) + \ldots + \lambda_j/(n+j) + x'/(n+j+1) + \lambda_j + 1/(n+j+2) + \ldots + \lambda_k/(n+k+1) + 0/(n+k+2) + \ldots \quad (1.19)$$

Notez que : $\forall i,\ j,\ \lambda_i \neq 0,\ \lambda_j \neq 0,\ i > j \Rightarrow \lambda_i \geq \lambda_j$

Alors, pour le calcul de $F_{A_r}\ \left(resp.\ F_{B_s}\ et\ F_{C_t}\right)$, nous prenons en compte la valeur x' (resp. y', z').

Pour calculer de $F_{A_rB_s}\ \left(resp.F_{A_rC_t},\ F_{B_sC_t}\ et\ F_{A_rB_sC_t}\right)$,nous prenons en compte la valeur $min\ (x', y')$ $(resp\ .min\ (x', z'),\ min\ (y', z'),\ min\ (x', y', z')$ .sachant que le nombre maximal de t_uplet pouvant être obtenus dans r_{su} est $na \times nb \times nc$, c'est-à-dire le produit des nombres d'étiquettes qui apparaissent dans les partitions . [30]

Exemple

Supposons, nous avons l'attribut: Nombre de grossesse de la base de données médicale PIMA Indiana Diabète. Nous calculons les cardinalités floues pour chaque degré d'appartenance de la variable linguistique '' Low '' afin de construire un résumé linguistique F_{low}

Donc : F_{low}=1.0/0+1.0/258+0.8/285+0.5/306+0.3/325

Notre résumé linguistique obtenu est interprétable au langage naturel. Alors interprétons le résultat cité ci dessus :

- nous avons 258 patients de la base de données PIMA qui appartiennent totalement au variable linguistique VL (low).
- nous avons 325-306=19 patients de la base de données PIMA qui appartiennent partialement au variable linguistique Vl (low) avec un degré=0.3.

Considérons que la cardinalité flou est un nombre graduel, noté par Nf.

De même que dans les nombres graduels, les opérations d'addition et de multiplication peuvent s'effectué dans l'espace de cardinalité flou.

Dans le résumé à base de calcul de cardinalité floue, pour chaque α-coupe, le comportement du nombre graduel est le même que celui d'un entier naturel. Ainsi pour réaliser l'addition de deux nombres graduels il suffit d'effectuer des additions pour chaque α-coupe.

Exemple A = {1/3, 0.7/4, 0.4/6} et B = {1/0, 0.9/1, 0.7/3, 0.5/4},

Pour l'addition, nous effectuons les calculs pour chaque coupe niveau(α-coupe) :

- ✓ La coupe de niveau 1 (le degré=1) :

Nous avons dans l'ensemble A (3 éléments) et dans l'ensemble B(0 élément).

Donc, $(A+b)_1 = 3+0 = 3$

- ✓ La coupe de niveau 0.9(le degré=0.9) :

Nous avons 3 éléments provenant de A (puisqu'ils sont présents à un degré supérieur) et 1 élément provenant de B .Donc $(A+b)_{0.9} = 3+1 = 4$.

- ✓ nous terminons jusqu'à niveau 0.4, on obtient le résultat :
  {1/3, 0.9/4, 0.7/7, 0.5/8, 0.4/10}.

Grace aux informations quantitatives et qualificatives obtenues par le résumé linguistique, à base de calcul de la cardinalité floue, nous pouvons répondre à des questions de type "Q éléments de la relation sont A" A un ensemble des variables floues portant sur des attributs, Q les quantificateurs flous (relatifs ou absolus).

1.7.2 Définition le Quantificateur flou

Un quantificateur flou est un sous-ensemble flou de E qui décrit un nombre de cas ou une approximative d'appartenance à l'ensemble E.

Les requêtes sont de la forme "une proportion Q de A de la relation sont B".

Donc, il faut déterminer la proportion floue d'éléments qui vérifient la propriété floue A, où la validité d'un résumé est définie par un degré qui exprime à quel point on est sûr que Q n-uplets de la relation vérifie la propriété floue A.[16]

Enfin, nous pouvons citer certains les avantages de résumé linguistique à base de calcul de la cardinalité floue :

- ✓ Si une étiquette n'est pas adaptée pour un attribut parce que peu d'éléments appartiennent à cette étiquette, alors, la meilleure valeur de réécriture est 0,4 . Par exemple dans $F_A = 1/0 + 0,4/1 + \ldots$
- ✓ Il est possible de répondre à des questions du type "Q éléments de la relation sont A" où A un ensemble des variables floues portant sur des attributs et Q les quantificateurs flou (relatif ou absolue).
- ✓ Cette technique de calcul de cardinalité floue apporte une utilisation des degrés d'appartenances qui sont étés absents dans le modèle SaintEtiQ.
- ✓ De plus de nombreuses pistes restent ouvertes quant à l'utilisation de la notion de la cardinalité floue.

1.8 Conclusion

Parmi les différentes méthodes existantes de réduction de données, nous traitons dans notre thèse le résumé linguistique à base de la cardinalité floue qui traite les données structuré de type attribut / valeur. Ce modèle utilise un ensemble de termes linguistiques, il décrit des données structurées sous une forme linguistique afin de produire des versions condensées des données. C'est à ce titre que les méthodes de résumés linguistiques s'inscrit dans l'un de meilleurs méthodes de la réduction sémantique des données.

Dans notre thèse de recherche, nous nous intéressons à développer des résumés linguistiques médicaux à base de calcul de la cardinalité floue. Cette technique de calcul de cardinalité floue nous apporte une utilisation des degrés d'appartenance qui était absente dans plusieurs méthodes.

Aussi, il y'a une possibilité d'utiliser les informations obtenues par ces résumés linguistiques pour exploiter un système d'interrogation flexible et de créer un nouveau type de classifieur supervisé médical.

chapitre 2. Interrogation flexible des résumés linguistiques

2.1 Introduction 27

2.2 L'interrogation de la Base De données relationnel 27

2.2.1 Définition de la base de données : 27

2.2.2 Système d'interrogation flexible 28

2.2.3 Requête floue : 30

2.3 Système d'interrogation des résumés linguistiques 31

2.3.1 Le modèle SaintEtiQ 32

2.3.2 Résumé de Yager 35

2.3.3 Approche de L.Liétard 39

2.3.4 Résume à base de calcul de Cardinalité floue 40

2.4 Conclusion 45

2.1 Introduction

Avec la démocratisation d'informatisation des données et leur utilisation comme un outil de communication, une somme de sources d'informations importantes est stockée sous une forme de la base de données. Face à cette multitude des données informatiques, il est important de développé un système de recherche efficace. Le but de ces systèmes de recherche est d'aider l'utilisateur de trouver l'information désirée parmi une masse de données disponibles.

Dans ces systèmes de recherche, l'utilisateur doit formuler son besoin de l'information sous la forme d'une requête où le rôle de ce système est d'identifier les enregistrements qui répondent à ces besoins. Ces systèmes fournissent une liste à l'utilisateur qui est triée en fonction de leur pertinence estimée (satisfaction). L'efficacité de ces systèmes est jugée par une meilleure adéquation entre la machine et l'humain. Si la représentation et la description des informations satisfont le besoins de l'utilisateur donc c'est un système d'interrogation flexible.

Le principe de Systèmes d'interrogation est de réduire à la fois l'espace de recherche et le nombre de sources interrogées (réduisant les temps de recherche), pour des réponses moins nombreuses et mieux ciblées (réduisant la surcharge d'informations).

Dans ce chapitre, nous allons tout d'abord rapidement présenter les systèmes d'interrogations de base de données traditionnels. Nous aborderons ensuite sur les ensembles flous, et leur application aux bases de données, notamment dans la gestion des requêtes plus flexibles. Nous essayons d'exploiter un état de l'art de différents algorithmes de recherche proposé pour chaque type de résumé linguistique et de montrer l'intérêt de ces résumés de données dans le contexte de système d'interrogation flexible.

2.2 L'interrogation de la Base de données relationnel

2.2.1 Définition de la base de données :

Une base de données est un ensemble structuré de données enregistrées dans un ordinateur et accessibles de façon sélective par plusieurs utilisateur [31]. Elle recueille des informations qui sont liées à un sujet donné. Une base de données peut être locale, c'est-à-dire utilisable sur une machine par un utilisateur, ou bien distribué, c'est-à-dire que les informations sont stockées sur des machines distantes et accessibles par le réseau. L'avantage de l'utilisation de bases de données est la possibilité d'être accédées par plusieurs d'utilisateurs simultanément.[32]

Mais après de nombreuse année de stockages des informations multimédia (Texte, images, Vidéo….) et le développement web sous forme de base de données, la recherche d'une information pertinente devient difficile sur un volume immense de données où l'utilisateur trouve une difficulté de la recherche d'une information pertinente dans un grand volume de données.

Face à cette multiplication des données informatiques, qu'elles soient structurées ou documentaires, la nécessiter d'un système de recherche efficace se fait de plus en plus sentir. Donc, le but de ces systèmes de recherche est d'aider un utilisateur à trouver l'information désirée parmi une masse de données disponibles.

Alors, il est nécessaire de trouver un moyen pour interroger ces données d'une façon rapide et optimale en utilisant le processus l'interrogation **LMD**, c'est-à-dire, nous faisons appel à la notion d'une requête.

Le processus d'interrogation LMD se compose en deux langages :

- **Le langage de programmation :** il se caractérise par un enchaînement d'instructions (itérations, conditionnelles, appel à des procédures ou des fonctions), d'affectation, de saisie, d'impression, de calcul d'expressions et de manipulation de structures de données élaborées.
- **Le langage d'interrogation (**constitué de deux sous langage) :
  - ✓ langage de désignation (**sélection, projection, jointure**)
  - ✓ langage de mise à jour (**insertion, modification, suppression**).[[32]

Nous pouvons dire que l'interroger une base de données permet de mettre des données à la disposition d'utilisateurs pour une consultation, une saisie ou bien une mise à jour, tout en s'assurant des droits accordés à ces derniers. L'utilisateur doit formuler son besoin sous la forme d'une requête, le système doit identifier les documents qui répondent à ces besoins [33].

Donc, la recherche d'une information pertinente est liée par le meilleur choix de la forme de la requête de l'utilisateur. Il faut que la requête formulé doive être compréhensible par le système informatique. Mais certains utilisateur par manque d'expérience, ne sait pas lui-même formulé clairement son besoin. Il existe des systèmes d'interrogations flexibles qui facilitent ces difficultés.

2.2.2 Système d'interrogation flexible

On a introduit, l'interrogation flexible est tend à rendre l'interrogation classique des BD plus souple parce que l'interrogation des bases de données (BD) nécessite une connaissance précise et détaillée des données et de leur organisation. Alors, l'efficacité du système d' interrogation flexible est lié par une meilleure adéquation entre la machine et l'humain [33].

Dans[34] [35], les Systèmes de Gestion de Bases de Données Relationnelles (SGBDR) sont devenus le noyau de tout système informatique. Ces recherches ont montré les limites des SGBDR sur le plan de l'interrogation flexible. Aussi, dans [36][37] ont été introduit une certaine forme de flexibilité dans l'interrogation des BD(Base de données).

En effet, l'interrogation classique d'une BDR est qualifiée d'interrogation booléenne dans la mesure où l'utilisateur formule une requête, avec SQL par exemple qui retourne un résultat

ou rien du tout. Cette interrogation pose des problèmes pour certaines applications [38]et [39]. L'utilisateur doit connaître tous les détails sur le schéma et sur les données de la BD. Plusieurs travaux ont été menés pour pallier à ce problème[40][41] .Nous trouvons aussi, plusieurs extensions du modèle relationnel et du langage SQL qui ont été proposées comme SQLf (L'interprétation des requêtes de l'utilisateur est basée sur la théorie des ensembles flous).Du façons général, nous déduisons que la création d'un système d'interrogation flexible efficace se fait à l'aide d'un système de personnalisation et de création une requête flexible.

2.2.2.1 Système de personnalisation

Un système de personnalisation permet d'avoir une utilisation spécifique(en fonction du profil de l'utilisateur), utilisé souvent sur le web. L'utilisation des préférences caractéristiques des utilisateurs se fait par l'ajout des informations et des contraintes. Ces informations permettent aux systèmes d'interrogations de réduire à la fois l'espace de recherche et le nombre de sources interrogées (réduction du temps de recherche), pour avoir des réponses mieux ciblées (réduisant la surcharge d'informations).[42]

2.2.2.2 Requête flexible

La requête flexible s'adapte aux systèmes d'interrogation afin de pouvoir interroger les données d'une manière flexible, c'est à dire les résultats peuvent être plus ou moins satisfaisants. Nous estimons, par le degré de satisfaction des conditions de la requête que les préférences de l'utilisateur peuvent s'appliquer aussi bien à des conditions élémentaires qu'à une combinaison de conditions.

En effet, le point clé de ces systèmes de recherche est l'évaluation de la pertinence des résultats vis-à-vis de la requête. L'information est souvent représentée sous forme de termes linguistiques. Ces termes peuvent provenir de documents textuels ou d'attributs linguistiques représentant des données numériques.

La notion de degré de satisfaction se compose par deux catégories :

- ✓ l'utilisation des ensembles flous pour représenter les termes imprécis des requêtes.
- ✓ l'extension spécifique du système relationnel (algèbre et langage).

Citons plusieurs systèmes d'interrogations flexibles existants :

- Les approches de[43][44] sont des approches orientées de recherche d'information. Ils utilisent la notion de distance par rapport à la valeur souhaitée.
- L'approche PREFERENCES de [45]: les résultats d'ensembles ou de listes de conditions booléennes sont agrégés numériquement, de manière à ce qu' un n_uplet soit pertinent et qui satisfait ces conditions, mais s'il ne satisfait pas une ces conditions , les suivantes serons ignorées.

- On a aussi l'approche [35] :le principe de cet approche est de reformuler la requête en termes d'ensembles flous .

Nous résumons que, la modélisation d'un système flexible sur des données classiques est de retrouver les données stockées dans la base de données qui sont défini par des requêtes flexibles.

Donc, le résultat de la requête n'est plus un ensemble d'éléments sélectionnés, mais un ensemble d'éléments discriminés en fonction de leur satisfaction globale.

L'une des solutions qui permet de s'approcher au mode de la pensée humain, caractérisé par l'imprécision et l'ambigüité, est l'utilisation de la logique floue, qui permet de modéliser les caractéristiques (requête) d'une manière formelle et de les intégrer dans un système d'évaluation.

2.2.3 Requête floue :

C'est une requête qui utilise la notion de la logique floue, elle permet de modéliser les caractéristiques de données d'une manière formelle, et de les intégrer dans un système d'évaluation. L'outil de construction de la requête floue se base sur la théorie des ensembles flous. Cette requête s'applique parfaitement aux données. Elle est proche du langage naturel.

La requête floue fait appelle à un filtrage de donnée quel que soit le domaine de la base de données ou de la recherche documentaire. Elle permet de représenter l'information qui satisfait un besoin. Ce dernier peut être exprimé d'une manière flexible (préférences de l'utilisateur). L'exemple typique de requête floue est : « **trouver les patients malades et le taux de GLU élevé**», décrivant une préférence sur l'état et le taux de Glu.

L'utilisation de la notion de logique flou permet de créer une requête plus flexible, pour cela considérons trois aspects importants de leur application :

- La formulation de requêtes flexibles doit s'adapter au modèle de la logique floue choisi, elle s'écrit par la pondération, les quantificateurs flous, les éléments de la requête et les operateurs d'agrégations adaptés.
- L'incertitude et l'incomplétude des données : La représentation des données doit être revue, en se basant sur les ensembles flous, pour définir des valeurs d'attributs mal connues.
- Le calcul de similarité des requêtes et des données : l'évaluation des requêtes sur ce type de données doit faire appel à des propriétés spécifique des modèles utilisés. Nous définissons les mesures de similarité ou de proximité entre les requêtes et les données pour voir la ressemblance entre les deux. [33]

Citons quelques travaux :

- Dans le travail proposé par [46] : Ils ont modélisé les processus d'agrégation des informations pondérées dans un cadre linguistique. Trois opérateurs d'agrégation d'informations linguistiques pondérées ont été présentés: opérateur de disjonction pondérée linguistique (LWD), opérateur de conjonction pondérée linguistique (LWC) et opérateur de pondération linguistique pondérée (LWA).
- Dans les travaux de[47] : Ils ont proposé une approche pour la sélection et le classement des services Web et ils ont réglé le problème des préférences de utilisateur par des prédicats flous. La satisfaction des préférences et la similitude structurale sont interprétées grâce à des quantificateurs linguistiques.
- Dans les travaux de [48], ils ont proposé un nouveau type de requête de base de données avec l'implication des préférences .l' interprétation de ces requêtes a été définie par la variante de l'ordre lexicographique et par une comparaison des distances minimales à une feuille de l'arbre condition.

En résumé que les prédicats flous permettent d'exprimer les préférences vagues de l'utilisateur, au moyen de propriétés voulues. Ils sont modélisés par des ensembles flous. Nous distinguons aussi que entre les éléments, il y'a une gradualité, car chaque prédicat a un degré d'appartenance.

Cette requête floue à deux avantages majeurs par rapport aux requêtes classiques. Elle permet :

- de donner des réponses approchées alors qu'une requête classique donne un résultat vide.
- de trier les bonnes réponses en fonction de leur degré de satisfaction, au lieu de donner une liste de réponses différentes.

Toujours, ce type de requête est défini par plusieurs constructeurs :

- les prédicats atomiques : tels que grand ou jeune, sont définis par un ensemble et sur un ou plusieurs domaines, par exemple, la taille et l'âge.
- les médicateurs : ils peuvent être appliqués aux fonctions d'appartenance qui définissent ces ensembles flous, afin de modéliser des variations linguistiques telles que (très, plutôt, etc). Ces médicateurs sont des fonctions du type dilatation, concentration ou translation [49]. Par exemple la fonction définie sur la taille est représenté par le prédicat (très grand), elle sera obtenue à partir de variable linguistique (grand), en effectuant une translation vers la taille croissante.
- Les expressions logiques : Les conditions précédentes peuvent être combinées, sous forme d'expressions logiques, par des opérations sur les ensembles flous (conjonctions et disjonctions). Le plus souvent, la conjonction est obtenue en appliquant l'operateur min sur les degrés, et la disjonction en appliquant l'operateur max. Aussi, les conjonctions et disjonctions peuvent être exprimées, en utilisant l'implication de Dienes, par [50].

2.3 Système d'interrogation des résumés linguistiques

Un résumé linguistique a pour un but de condenser les informations sémantiquement ; son principe est d'utiliser la notion de la logique floue. En général, la logique floue nous offre un

résumé intelligent où le résumé est associé par des degrés d'appartenances et des degrés de satisfactions.

Donc, un résumé linguistique est interprétable en langage naturel parce qu'il utilise, pour la présentation, les variable linguistique de la logique flou. Cette présentation est proche du langage naturel, ainsi que la recherche d'une information à l'aide d'une requête. Les résumés de données sont utilisés pour améliorer l'efficacité du processus d'interrogation. L'exploitation de ces résumés, nous permet de tirer plus de connaissance que l'interprétation des résumés. Elle doit notamment offrir à l'utilisateur la possibilité de préciser les critères de regroupement qui semblent intéressants pour un traitement particulier.

Le système d'interrogation d'un résumé linguistique est considéré comme un moyens de répondre au problème de projection par ce que l'utilisation de résumes accélère l'accès aux données et conduit à des réponses plus compactes, et vagues.

Ces systèmes permettent de répondre à des requêtes telle que : « **Comment les individus sont x »** Où **X** est une projection de la relation **R** sur un ensemble ordonné d'attributs défini dans la requête ($X \subseteq R$). Une telle requête spécifie les individus ou les objets qui prennent en compte l'indication des caractéristiques x portant sur les attributs **X**. Donc, répondre à une telle requête, consiste à faire du prétraitement dans les résumés afin de trouver des résultats.

L'interrogation n'a plus pour but de décrire les données mais elle est considérée comme un test d'existence. Elle peut savoir si des données présentant des caractéristiques spécifiées existent ou non. L'interrogation des résumés fournit une réponse à tous les types de requêtes.[51].

Dans les différentes méthodes de résumé linguistique, il est possible d'utiliser la requête flou dans d'un système d'interrogation flexible. Mais chaque méthode de résumé linguistique à son propre traitement de recherche de l'information qui convient à la demande d'utilisateur.

Nous citons par la suite un état de l'art sur les différents algorithmes du processus d'interrogation pour chaque méthode de résumé linguistique.

2.3.1 Le modèle SaintEtiQ

Le modèle SaintEtiQ est une méthode de construction de résumés linguistique de données qui sont structurées et composées de couples attribut/valeur. [51]

Dans cette approche, Les résumés de données sont utilisés pour améliorer l'efficacité du processus d'interrogation où les utilisateurs peuvent définir et utiliser leur propre vocabulaire (au lieu du domaine des attributs). Ce vocabulaire est défini à l'aide de variables

linguistiques, stockées dans le profil de l'utilisateur, et permet de définir la granularité souhaitée, aussi bien dans les requêtes que dans les réponses.

il y'a plusieurs d'algorithmes d'interrogation de résumés SAINTETIQ citons :

- Dans les travaux de [19] ils sont les premiers qui ont proposé l'algorithmes de recherche d'une base de données ,utilisant le vocabulaire prédéterminé des résumés de SaintEtiQ .
- Dans les travaux de(W A Voglozin, Raschia, Ughetto, & Mouaddib, n.d.,2004) [20] ils ont proposé une approche qui étend un travail antérieur sur l'interrogation flexible, et les outils d'utilisation des résumés.
- Dans les travaux de[[53][20]]. Ont proposé des améliorations sur le premier d'algorithme d'interrogation flexibles pour les bases de données qui traites des résumés .
- Dans les travaux de[20] .Ont proposé un algorithme qui peut être considéré à la fois comme un mécanisme de requête booléen sur une hiérarchie de résumés et comme un mécanisme d'interrogation flexible sur les uplets relationnels sous-jacents .
- [54][52]Les algorithmes proposés ici étendent ceux à l'utilisation du vocabulaire spécifique à l'utilisateur.
- De même, dans[54], Ils ont essayé de traité des algorithmes d'interrogation manquants .

2.3.1.1 La Forme général de la requête :

Le modèle SaintEtiQ permet de répondre à plusieurs questions simples qui portent sur l'ensemble des attributs, n'intervenant pas dans les critères de sélection.

La formulation d'une question s'écrit :

Comment sont sur Y les individus qui sont x sur X ? (2.1)

Où X est l'ensemble des critères de sélection, et Y le complémentaire de X sur l'ensemble de la relation.

Dans toute requêtes telle que (2.1) où X définit implicitement une relation complémentaire Y par rapport à l'ensemble A des attributs de R, les attributs de X sont considéré comme des attributs d'entrés et Y comme des attributs de sortis .

Chaque valeur d'attribut $x.A_i$ explicite les caractères requis sur l'attribut A_i . Le champ $x.A_i$ est multi values, c'est-à-dire composée de plusieurs valeurs.[56][57]

Exemple : Comment sont les rouleaux malléables ?

Il s'agit de décrire des rouleaux par leur épaisseur et leur température de fusion.

On a **X = (dureté)** et **Y = (épaisseur, température).** Il n'y a qu'un caractère requis (sur l'attribut dureté) : **C1 = Cdu = {doux}** et la caractérisation initiale est **C = { {doux} }.**

Le langage d'interrogation utilisé, est un langage de requête proche de **SQL** qui commence par le mot clef « **DESCRIBE »**.

La forme de la requête est, pour le reste, très proche d'une sélection traditionnelle en SQL , suite à cela ,la requête prend la structure suivante **:**

DESCRIBE [<table>] ON <liste_d_attribut> WHERE <condition>.

La forme générale d'une requête est :

DESCRIBE [<table>] ON <liste d'attributs> WHERE <conditions de sélection> [57]

2.3.1.2 Évaluation des requêtes :

Comme le rappelle Chris Date dans[58] , l'évaluation d'une requête dans un système de gestion des bases de données comprend habituellement quatre étapes :

- ✓ l'expression de la requête dans un format interne.
- ✓ la réduction de l'expression interne à une forme canonique.
- ✓ le choix des procédures d'exécution.
- ✓ la génération de plans de requête et le choix du meilleur plan.

Dans le cadre de l'interrogation des résumés linguistique, le format interne est représenté par la notation ensembliste des caractérisations. Nous pouvons considérer qu'une forme canonique est atteinte puisqu'il n'y a qu'une représentation en extension d'un ensemble fini. L'existence d'une seule opération dans l'interrogation et la caractérisation rend inutile les deux dernières étapes.

En général, les résumés produits par SaintEtiq sont stockés dans des fichiers XML. Ils suivent un ordre préfixé et typique de la sérialisation de données arborescentes. Le parcours en profondeur permet de réaliser des coupures de branches qui contribuent largement à réduire le nombre de nœuds visités pendant la recherche. [2]

Alors pour évaluer une requête, le système parcourt l'arbre de la hiérarchie de résumés à partir de sa racine. A chaque nœud de la hiérarchie un processus, le système de recherche est lancé afin de déterminer , il faut continuer l'exploration des sous-arbres ou arrêter l'exploration.

Voici les différents cas qui peuvent être rencontré lors de l'exploration :

- Cas 1, arrêt de l'exploration par défaut de réponse : c'est le cas où il existe une clause (ou plus) pour laquelle le résumé ne satisfait pas le critère de sélection
- Cas 2, arrêt de l'exploration par satisfaction de la requête : pour chaque attribut présent dans la requête, le résumé étudié présente uniquement les mêmes critères.

Pour chaque attribut, l'ensemble des étiquettes, présentes dans le résumé, est inclus dans l'ensemble des étiquettes de la requête.

- Cas 3, poursuite de l'exploration : si pour au moins un attribut de la requête, des étiquettes sont dans le résumé et dans la requête, alors il est possible qu'il y a des éléments à exclure de ce résumé pour répondre à la question.

Exemple : Si la requête était :

Comment sont les rouleaux d'épaisseur moyenne ?

On aurait :

- **X = (épaisseur) • Y = (dureté, température)**
- **C = {{moyen}}**

• P = moyen Et si les mêmes résumés étaient sélectionnés, c'est-à-dire **z1, . . . , z6**, on aura comme seul résultat **P** (puisque **P** n'a qu'une interprétation) :

• C = { {doux, mou}, {modéré, normale, bas} }.

Afin de faciliter la lecture des résultats, les résumés présentant les mêmes ensembles de descripteurs pour tous les attributs requis peuvent fournir une liste de descripteurs par attribut :

• fin, mou ⇒froid ; • fin, doux ⇒froid, bas.

Les requêtes qui sont réalisables avec SaintEtiQ sont simples mais limitées. Ce modèle ne permet pas l'interrogation sur des quantités d'éléments, présenté dans un résumé, par une étiquette linguistique[57]. Il ne permet pas la résolution à des questions importantes dans le domaine médical, tel que **« Combien de patients sont jeune ?** ».

Donc, le modèle SaintEtiQ ne permet pas de répondre à certaines questions qui paraissent pourtant intéressantes à résoudre.

2.3.2 Résumé de Yager

Le résumé linguistique quantifié de données proposée par Yager [3]] utilise la notion de théorie de la logique floue sous une forme simple . Il utilise les variables linguistiques des ensembles de données. C'est la plus ancienne méthode utilisée par les développeurs malgré les inconvénients, leur principe de construction de résumé est basé sur recherche et la validation d'une requête (chapitre1 section 1.6).

Tous d'abord, la forme général de résumés de données linguistiques, proposé par Yager, est illustrée par : «la plupart des patients sont jeunes et Glu élevé» (avec un certain degré de vérité) et leurs extensions. Il utilise le concept d'une protoforme (forme prototypique), qui a

été proposé par Zadeh. La protoforme est considérée comme une forme générale d'un résumé de données linguistiques. [59]

Le système d'interrogation du résumé de Yager est basé sur les informations de résumé linguistique , qu'il utilise la notion de la logique floue et des requêtes de bases de données floues.

Il existe plusieurs travaux traitant l'algorithme d'amélioration d'un système d'interrogation flexible de Yager:

- Dans les travaux de[60] [61],ils ont proposé quelques extensions à la forme d'un résumé linguistique.
- Dans [62] ,ils ont proposé une nouvelle forme qui s'inscrit toujours dans le schéma d'une règle d'association. Par rapport aux approches précédentes , [63] ont proposé un algorithme pour extraire les règles d'association via le paquet **FQUERY for Access** de l'auteur et ils ont montré que l'idée de Zadeh de protoforme, et leurs hiérarchies, peuvent être employées pour représenter divers types de résumés linguistiques.
- Dans [60], ils ont présenté une autre extension de l'approche classique de Yager ,interactive aux résumés linguistiques flous, basés sur la logique floue et les requêtes de bases de données floues. Ils ont montré comment les requêtes floues sont liées à des résumés linguistiques, et comment on peut introduire une hiérarchie de protoforme.
- Dans [64],ils ont étendu une approche de résumé linguistique des séries temporelles (numériques) en proposant d'utiliser les opérateurs OWA (ordonnancés à moyenne pondérée) pour l'agrégation des tendances partielles .
- Dans [24],il ont montré qu' un résumé des données linguistiques d'un ensemble de données est souhaitable et cohérent pour tous les services du personnel. Pour extraire des résumés de données linguistiques complexes, l'opérateur LOWA est utilisé à partir de la logique floue.

2.3.2.1 Langage d'interrogation de résumé Yager

Langage utilisé dans le model proposé par Yager est appelé **SummarySQL** [23].le langage **SummarySQ**L a été développé par Rasmussen et Yager (1997).C'est un langage de requêtes floues destiné à intégrer des résumés dans une requête floue. Le langage peut évaluer le degré de vérité d'un résumé proposé par l'utilisateur. Il peut également utiliser un résumé comme prédicat dans une requête floue. Un résumé exprime la connaissance de la base de données dans un état sous la forme «**Q objects in BD areS**» ou «**QRobjects in BD areS**» .où Q est un quantificateur linguistique et R et S sont des résumés (termes linguistiques).

Par exemple: «**la plupart des gens dans BD sont grands**» ou «**la plupart des personnes de grande taille dans BD sont lourds**». Les prédicats et les termes linguistiques sont des ensembles flous dans l'expression qui représente la condition de sélection. Cette expression est évaluée pour chaque t_uplet des valeurs d'appartenance associées qui sont ensuite

utilisées pour obtenir une valeur de vérité pour le résumé. **SummarySQL** est utilisé pour déterminer dans quelle mesure une instruction est vraie. Ce langage peut également être utilisé pour rechercher des règles floues.

Donc, Ce modèle permet d'interroger une relation pour poser des questions du type "**Q éléments de la relation sont A**" avec Q un quantificateur, tel que "la plupart", et A une étiquette linguistique pour un attribut V.

Le calcul s'effectue suivant un algorithme définit dans [27] comme suit :

Pour chaque n-uplets O_i de la relation, nous calculons de degré d'appartenance de la valeur de l'attribut V à l'étiquette A. Cette valeur est notée $\mu_A(O_i)$ (car µA est la fonction caractéristique de A). Ensuite une fois cette valeur calculée pour chaque n-uplets , on évalue la véracité moyenne **'r'** par la formule :

$$r = \frac{1}{n}\sum \mu_A(O_i) \qquad (2.2)$$

Enfin on calcule la validité du résumé par rapport au quantificateur avec la

formule : $t = \mu_Q(r)$.

Toutefois si la requête est de la forme "**Q B éléments de la relation sont A**", le calcul de la vérité moyenne diffère. En effet il est nécessaire de prendre en compte la présence de l'étiquette B. Le calcul de la vérité est alors donné par la formule :

$$r = \frac{\sum norm(\mu_B(O_i), \mu_A(O_i))}{\sum \mu_B(O_i)} \qquad (2.3)$$

Ces différents calculs ont été automatisés afin de permettre une interrogation directe d'une base de données par l'intermédiaire d'un langage proche de SQL. La syntaxe définit dans [YaRa97] pour une sélection simple d'élément d'une table est défini par :

```
SELECT attributs
FROM tables
WHERE conditions
```

Avec la possibilité de placer des étiquettes linguistiques dans la condition. Cette sélection permet d'obtenir l'ensemble des n-uplets qui vérifient la condition de degré de la correspondance de chaque n-uplets. Il est aussi possible de soumettre des résumés afin de les évaluer. Ceci se fait à l'aide de la structure :

```
SUMMARY quantificateur
FROM tables
WHERE conditions
```

La découverte des règles floue est basée sur l'hypothèse qu'on peut interpréter une requête du type **"Si un n-uplet est B alors il est A"** par "**presque tous les B sont A**".

Donc, le langage **SummarySQL** présente une technique de résumé qui est basé sur le calcul de l'appartenance d'un attribut et le comptage des valeurs d'appartenance de chaque n-uplet , pour obtenir un degré de vérité d'un résumé [16].

Il existe des autres propositions de langage pour l'interrogation comme :

- **SQLF**

  Le langage de requête SQLf, proposé par[36], est une extension de SQL visant à introduire les prédicats flous dans SQL autant que possible. Une augmentation de la syntaxe et de la sémantique de SQL sont effectuées de sorte que la plupart des éléments d'une requête peuvent être fuzzifier. Les éléments de SQlf comprennent des opérateurs, des fonctions d'agrégation, des modificateurs (très, vraiment, plus ou moins), des quantificateurs (la plupart, une douzaine) ainsi que des termes de description générale (jeune ou GLU élevé). Un exemple de requête dans SQLf est : **"select 10 service du groupe PATIENTS ayant la plupart des (âge = jeune) sont Glu élevé"** où les mots-clés SQL standard sont en gras, **service** et **âge** sont des attributs d'une relation nommée **PATIENT**S. La requête sélectionne les 10 service qui ont le plus de satisfaction de la condition «la plupart des jeunes patients sont Glu élevé».[65]

- **Fquery**

  Le langage FQUERY [55]est une intégration de l'interrogation flexible dans un système de gestion de base de données. Le système permet des requêtes avec des prédicats flous qui sont exprimés par des ensembles flous. Les requêtes peuvent contenir des quantificateurs linguistiques et permettent d'attribuer différents niveaux d'importance aux attributs. Les auteurs tentent d'appliquer le paradigme de l'informatique et, éventuellement, de traiter les valeurs linguistiques, les quantificateurs, les modificateurs et les relations.

  FQUERY utilise des ensembles flous pour l'aspect d'imprécision et exécute une syntaxe et une extension sémantique de SQL. Les valeurs linguistiques et les quantificateurs sont représentés sous la forme d'ensembles flous. Du côté de la sémantique, la requête est considérée comme un ensemble flou résultant de la combinaison d'ensembles flous des valeurs linguistiques et de quantificateurs. Par conséquent, chaque enregistrement sélectionné par une requête SQL classique a un degré de satisfaction, utilisé dans une étape de classement, puisqu'il indique à quel point l'enregistrement correspond à la requête.

Nous résumons que l'utilisation du résumé de Yager dans un système d'interrogation flexible, ne nous donne pas toujours des bons résultats, le résumé de Yager reste incomplet parce qu'il faut valider tous les cas d'une requête ,afin de construire une base de connaissance complète. Alors, nous pouvons dire que ce n'est pas un vrais résumé, par ce qu'il des fois utilise plus espace que la base de données numérique. Les algorithmes d'améliorations proposés par les chercheurs, utilise toujours la base de données numérique. En plus, la validation d'une requête est basée sur le calcul de degré de vérité où la recherche proposé par Liétard a prouvé que le calcul de degré de vérité est insuffisant pour valider une requête.

2.3.3 Approche de L.Liétard

Au début, la méthode proposée par Yager utilise la degré de vérité pour vérifier la validité d' une requête floue, il a utilisé comme forme de la requête ,la protoforme proposer par Zadeh(QBX are A).

$$\gamma = \mu_Q\left(\frac{\sum coun(XA \cap XB)}{\sum count(XB)}\right) \qquad (2.4)$$

Où XA (resp.XB) est un ensemble flou constitué d'éléments de X qui satisfait à la condition A (respectivement B),mais ,ce résumé n'a que les informations quantitatives.
Alors, la proposition de Yager ne montre pas suffisamment la relation entre les deux aspects (les informations quantitatives et les informations qualificatives) .

Dans le travail de L.Liétard[5], il est proposé un nouveau type de requête linguistiques . Cette requête est faite de déclaration linguistique associée à un degré de validité (une nouvelle définition du résumé linguistique).

L'algorithme de cette approche est :
Soit $C^i, C^{i+1}, ..., C^{i+k}$ les prédicats flous respectivement définies sur les domaines des attributs $A^i, A^{i+1}, ..., A^{i+k}$ de la relation R et les requêtes linguistiques de type:

<<typle from R satisfy C^i and C^{i+1} and…and C^{i+k} >>

Condition << C^i and C^{i+1} and…and C^{i+k} >> est une contrainte floue exprimée par le résumé. Le degré de validité de tel résumé est donné par:

$$\omega = \max_{\alpha\in[0,1]} \min\left(\alpha, F\left(C_\alpha^i \times ... \times C_\alpha^{i+k}\right)\right) \qquad (2.5)$$

Où F est une fonction définir par:

$$F(E) = \left|E \cap R\left[A^i, ..., A^{i+k}\right]\right| / \left|R\left[A^i, ..., A^{i+1}\right]\right| \qquad (2.6)$$

Dans ce travail, il suffit de calculer le degré de validité pour chercher le quantificateur flou d'une requête complexe.

2.3.4 Résume linguistique à base de calcul de Cardinalité floue

2.3.4.1 La définition de résumé linguistique à base de calcul de la cardinalité floue

L'approche de résumé linguistique à base de cardinalité floue, présentée dans[4][18],se développe en deux parties.

- La première étape, on commence par la phase d'étiquetage
- La deuxième étape, on commence par le travail de calcul de cardinalités floues.

Il y'a plusieurs représentations de résumé à base de calcul de cardinalité flou qui ont été proposées. Nous allons nous intéresser ici à la méthode proposé par D. Rocacher dans[29] [66]. Elle consiste à représenter la cardinalité d'un ensemble flou par un nombre graduel.

Le nombre graduel est basé sur la définition de la cardinalité floue, définie par :

$$\forall n \in R, \mu_{card(E)}(n) = \sup\left\{\alpha \middle| \ddot{}card(E_\alpha) \geq n\right\} \qquad (2.7)$$

Où :

- ✓ Le degré α est compris entre 0 et 1 (coupe)
- ✓ E un ensemble flou.

La cardinalité floue est définie pour un attribut A par :

$$FA = 0/0 + \ldots + 0/(n-1) + 1/n + \lambda 1/(n+1) + \ldots + \lambda k/(n+k) + 0/(n+k+1) + \ldots \qquad (2.8)$$

Avec $1 > \lambda 1 \geq \ldots \geq \lambda k > 0$.

Par exemple, $A = \{1/x_1,\ 1/x_2,\ 0.8/x_3,\ 0.7/x_4,\ 0.7/x_5,\ 0.5/x_6\}$.

$\mathrm{card}(\mathrm{A}) = \{1/0,\ 1/1,\ 1/2,\ 0.8/3,\ 0.7/4,\ 0.7/5,\ 0.5/6$. On est sûr que A contient totalement deux éléments. 0,8 représente dans quelle mesure A contient 3 éléments. Cette représentation de la cardinalité permet de connaître ,pour chaque coupure de niveau (**α-coupe),** le nombre d'éléments de l'ensemble A qui ont un degré supérieur à α où **α** est défini sur l'intervalle **[0,1]**.[57]

Avec les cardinalités floues calculés, dans le résumé linguistique, nous pouvons dire qu'il est possible de répondre à des questions du type "Q **éléments de la relation sont A**" avec **A** un ensemble d'étiquettes floues portant sur des attributs d'entrées et **Q** un quantificateur flou. En pratique, les quantificateurs sont plus intéressants, du point de vue des résumés, comme le quantificateur relatifs par exemple, la plupart. Nous effectuons des calculs sur les étiquettes de requête où nous utilisons les opérateurs tels que l'addition ou la multiplication sur informations de l'étiquette.

Pour chaque α-coupe, le comportement du nombre graduel est le même que celui d'un entier naturel. Ainsi pour réaliser l'addition de deux nombres graduels, il suffit d'effectuer des additions pour chaque α-coupe.

Exemple A = {1/3, 0.7/4, 0.4/6} et B = {1/0, 0.9/1, 0.7/3, 0.5/4},

Pour l'addition, on effectue le calcul pour chaque coupe niveau :

- ✓ La coupe de niveau 1 (le degré=1) :

  On a dans l'ensemble A (3 éléments) et l'ensemble B(0 élément).

Donc, $\left(A+b\right)_1 = 3+0 = 3$

- ✓ La coupe de niveau 0.9(le degré=0.9) :

On a 3 éléments provenant de A (puisqu'ils sont présents à un degré supérieur) et 1 élément provenant de B .Donc $\left(A+b\right)_{0.9} = 3+1 = 4$.

- ✓ On termine jusqu'à niveau 0.4, on obtient le résultat :
  {1/3, 0.9/4, 0.7/7, 0.5/8, 0.4/10}[57]

Les exemples des requêtes qu'on peut traiter :

- calculer la moyenne des salaires des employés prés de la retraite;
- trouver les entreprises où le nombre des employés sont jeunes
- trouver les entreprises dont la plupart des employés jeunes sont bien-payés.

Nous savons que l'interrogation flexible permet de prendre en compte des préférences de l'utilisateur dans les requêtes .Alors, grâce aux informations qui sont présentées par le cardinalité flou et leurs degrés d'appartenances, nous pouvons traiter des requêtes qui contiennent des prédicats quantitatifs .Elles sont des nombres graduels et calculables. les calculs qui s'effectue sur les notions de préférence et de cardinalité, nous conduits à définir le concept d'entier graduel ($\mathrm{N}f$).

D'après [29] le calcul entre les quantités graduels , nous permettent de traiter des requêtes flexibles complexes et des requêtes de plusieurs types .nous détaillerons ces différents types de quantité graduels ci-dessus.

2.3.4.2 Quantité graduel

La théorie de l'ensemble flou est la meilleure méthode qui peut exprimer des requêtes flexibles. Il a été montré dans[67][43][68] que calculs d'une requête flexible est basé sur des distances.

En effet, un ensemble flou E est défini par une fonction caractéristique $\mu_E(x)$ appartenant à l'intervalle [0, 1], qui exprime dans quelle mesure l'élément x appartient à l'ensemble flou E.

Les prédicats d'une requête sont graduels si le résultat a un degré de satisfaction, par exemple : jeune et bien-payé, sont décrits au moyen d'ensembles flous. Nous pouvons combiner les prédicats (critère) grâce aux opérateurs de l' Entiers naturel graduel (conjonction ou de disjonction ou de moyennes) [66].

2.3.4.2.1 Entiers naturel graduel :

Un ensemble flou, défini sur un domaine X, la fonction caractéristique est notée par : μ_E, telle que :

$$\mu_E : X \to [0, 1]$$

$$x \to \mu_E(x). \qquad (2.9)$$

Où la valeur $\mu_E(x)$ exprime dans quelle mesure l'élément x de X appartient à l'ensemble flou E.

- ✓ Si $\mu_E(x)=0$, Alors x n'appartient pas à l'ensemble flou E
- ✓ si $\mu_E(x)= 1$, alors x est appartient complètement dans E.
- ✓ si $\mu_E(x)$ est proche de 1 (resp. 0), plus (resp. moins) x appartient à E.

Dans le cas d'un ensemble flou fini E, on note par :

$$E = \left\{ \mu_E(x_1)/x_1, \ldots, \mu_E(x_n)/x_n \right\} = \sum_{i=1}^{n} \mu_E(x_i)/x_i \qquad (2.10)$$

Grâce à la notion de *coupes de niveau* ou α-*coupes*, un ensemble flou peut être décrit comme une collection d'ensembles ordinaires.

La coupe de niveau α de l'ensemble flou E, notée E_α, est l'ensemble composé des éléments dont le degré d'appartenance à E est au moins égal à α, d' où :

$$E_\alpha = \left\{ x / x \in X et \mu_E(x) \geq \alpha \right\} \qquad (2.11)$$

La cardinalité floue $|E|$ d'un ensemble flou E est définie par un ensemble flou d'entiers caractérisé par :

$$\forall\, n \in , \mu_{|E|}(n) = sup\{\alpha \,/\, |E_\alpha| \geq n\}. \qquad (2.12)$$

Où le degré α, associé à un entier n de $|E|$, évalue dans quelle mesure E contient au moins n éléments.

Exemple :

Soit l'ensemble flou E = {1/x_1, 1/x_2, 0.8/x_3, 0.4/x_4}, la cardinalité de E est représentée par :

$|E|$ = {1/0, 1/1, 1/2, 0.8/3, 0.4/4}.

Le degré 0.8 de $|E|$ exprime dans quelle mesure l'ensemble flou E contient au moins 3 éléments. [69]

Alors selon [70]la cardinalité d'un ensemble usuel fini qui est considéré comme un entier naturel n .Dans l'ensemble de α-coupe, un entier graduel x est un ensemble d'entiers formant une suite croissante {0, 1, …,$x\alpha$} qui peut être représentée par sa plus grande valeur $x\alpha$.

On appelle coupe de niveau α d' un entier graduel x, le plus grand entier de l'α_coupe de l'ensemble flou d'entiers défini par x. Ce plus grand entier de l'α_coupe est interprété comme un entier positif $x\alpha$ où nous pouvons effectuer cette opération sur un multi ensembles flous [70].

2.3.4.2.2 Opération sur un multi ensemble flou :

En effet, un élément x, dans un multi-ensemble flou A(*i.e.* multi-ensemble dont les occurrences des différents éléments sont associées à un degré d' appartenance), est caractérisé par un entier graduel, noté $\Omega_A(x)$, représentant la cardinalité de l' ensemble flou de des différents occurrences de x dans A. Les opérations utilisées dans les multi-ensembles flous sont :

- Intersection, union, union additive, produit cartésien, …
- les opérations de base sur les entiers graduels (min, max, $\oplus$, $\otimes$, …).
- l'opération binaire.

La définition d'une opération binaire * entre deux entiers graduels Q et Q' est une extension étendu qui est définie par :

$$\mu_{Q*Q'}(z) = \sup_{(x,y)/x*y\geq 0} \min(\mu_Q(x), \mu_{Q'}(y)) \qquad (2.13)$$

Les opérations * sur les entiers graduels respecte la propriété caractéristique :

$$(x*y)_\alpha = x_\alpha * y_\alpha \qquad (2.14)$$

Exemple : On a multi-ensemble flou : A = {<1, 0.1,0.1>/a, <0.5>/b} il est signifie que A contient trois occurrences de l'élément a, chacune étant affectée d'un degré d'appartenance, respectivement 1, 0.1et 0.1, et une occurrence de b au degré 0.5.

Les nombres graduels d'occurrences des éléments sont :

$\Omega_A(a)$ = {1/0, 1/1, 0.1/2, 0.1/3} ;

$\Omega_A(b)$ = {1/0, 0.5/1}.

A peut également se noter : A = {{1/0, 1/1, 0.1/2, 0.1/3}*a, {1/0, 0.5/1}*b}.

Si B est le multi-ensemble flou {{1/0, 1/1, 0.5/2}*a, {1/0, 0.5/1}*b}.

L'union additive $A + B$, obtenue en regroupant les éléments de A et de B, est définie par :

$\Omega_{A+B}(a) = \Omega_A(a) \oplus \Omega_B(a)$ = {1/0, 1/1, 0.1/2, 0.1/3}

$\oplus$ {1/0, 1/1, 0.5/2} = {1/0, 1/1, 1/2, 0.5/3, 0.1/4, 0.1/5}

Ω_{A+B}(b) = $\Omega_A(b) \oplus \Omega_B(b)$ = {1/0, 0.5/1} $\oplus$ {1/0, 0.5/1} = {1/0, 0.5/1, 0.5/2} [66]

Nous résumons que l'intérêt de cette démarche est d'offrir une base algébrique permettant la composition de calculs, et cette approche a été étendu au nombre relatif Zf et au rationnel Qf et leurs opérateurs (addition, soustraction, division, multiplication) ont été définis. Ces contextes permettent de traiter des requêtes flexibles complexes basées sur des calculs entre les quantités graduelles, et d'étudier des conditions flexibles telles que : quantificateurs absolus (au moins 3, environ 5), relatifs (la plupart, environ la moitié, presque tous, peu de) .

2.3.4.2.3 Calcul graduel sur la requête flou.

L'utilisation des calculs de nombre graduel dans le résumé à base de calcul de cardinalité floue, nous permet de savoir combien de t_uplet sont présents dans un résumé. Il suffit de déterminer la cardinalité floue de l'ensemble flou des t_uplet associé à un nœud. Une telle cardinalité peut se représenter par un nombre graduel.

Avec ce modèle, on peut poser facilement des questions du type : « **Combien de personnes sont de taille grande ou moyenne à un degré supérieur à 0,7 ?** ». C'est avec ce genre de question que le modèle prend son intérêt.

En effet, peu de modèle offre la possibilité de pouvoir sélectionner les t_uplets en vérifiant un critère à un certain niveau de satisfaction.

L'analyse de la structure a permis de mettre en évidence que, avec l'utilisation de cardinalité graduelle, de nombreuses informations n'ont pas besoin d'être pré-calculées mais peuvent être calculées de manière dynamique. En effet, les cardinalités graduelles permettent de calculer la cardinalité d'une union en fonction des ensembles et de leur intersection.

On obtient donc la formule :

$$card(A \cup B) = card(A) + card(B) - card(A \cap B) \qquad (2.15)$$

Où A et B sont des sous-ensembles flous.

La démonstration se fait par α-coupe et s'appuie, d'une part, sur une propriété des α-coupe : $(A \cup B)_\alpha = A_\alpha \cup B_\alpha$ et d'autre part sur une propriété des cardinalités d'ensemble ordinaire.

Alors :

$$card\left(\left(A\cup B\right)_{\alpha}\right) = card\left(A_{\alpha}\cup B_{\alpha}\right) = card\left(A_{\alpha}\right) + card\left(B_{\alpha}\right) - card\left(A_{\alpha}\cap B_{\alpha}\right) = card\left(A_{\alpha}\right) + card\left(B_{\alpha}\right) - card\left(\left(A\cap B\right)_{\alpha}\right). \quad (2.16)$$

Sachant que $card\ (A), \mathrm{card}(b), \mathrm{card}(A\cap b)$ est enregistré dans le modèle de résumé linguistique.

Même, Il y'a d'autre forme de requête qui peut être réalisé à l'aide des informations de résumé linguistique à base de calcul de la cardinalité floue, nous montrons ci-dessous les différentes propositions effectué afin d'améliorer le système d'interrogations flexible sur ce résumé linguistique :

- Dans les travaux de [48],l'approche des préférences hiérarchiques floues, proposée, utilise des degrés de préférence dans l'intervalle unitaire et combine les degrés d'appartenance au moyen d'une agrégation min-max.
- Dans [71] on a étudié comment un modèle basé sur probabilités de quantification floue peut être utilisé pour construire des résumés flous quantifiés, et comment construire un ensemble d'expressions quantifiées floues pour résumer les données. De plus, on a traité des expressions quantitatives proportionnelles unaires.

Cette technique de calcul de cardinalité apporte une utilisation des degrés d'appartenance qui était absente dans le modèle SaintEtiQ. De plus nombreuses pistes restent ouvertes quant à la possibilité d'utilisé la notion de cardinalités floues pour ce qui est de l'évaluation[16].

2.4 Conclusion

Ce chapitre a abordé l'exploitation de résumé linguistique dans le système l'interrogation flexible. Au début, nous citons les différents algorithmes proposé par les chercheurs afin de lier les résumés linguistiques avec la requête floue, afin d'obtenir un système de recherche rapide et qui satisfait les besoins de l'utilisateur. Nous citons ainsi les avantages et les inconvénients pour chaque système de recherche.

Pour cela dans notre thèse , nous proposons d'exploiter le résumé linguistique à base de calcul de la cardinalité floue afin de créer un système interrogation médical qui répond au besoin d'utilisateur (médecin) en utilisant la requête floue.

Notons qu'il y'a une infinité de quantificateurs flous. Les différents travaux qui ont développé les systèmes d'interrogations de résumé linguistique à base de calcul de cardinalité floue, sont concentrés sur le problème des quantifieurs flous de la requête.

Notre contributions est utilisé le principe d'approche de Liétard [28][72]afin de créer, dans le domaine médical, un système d'interrogation sur le résumé linguistique à base de calcul de la cardinalité floue.

chapitre 3. La relation entre le résumé linguistique et la classification supervisée

3.1 Introduction 47

3.2 Extraction de connaissance à partir de données 47

3.2.1 Définition de Data Mining (Fouille de données) 47

3.3 La classification 48

3.3.1 L'apprentissage 49

3.3.2 Principe de la classification 50

3.3.3 Classification supervisée 51

3.4 Les méthodes de classification 52

3.4.1 Séparateurs à Vaste Marge 52

3.4.2 Arbre de décision 54

3.4.3 Réseaux de neurone 55

3.4.4 Les plus proches voisins 57

3.5 Prédiction et le calcul de la distance métrique des résumés linguistique 60

3.5.1 Présentation de résumé linguistique 60

3.5.2 Evaluation de la similarité d'une protoforme classique : 63

3.5.3 Similarité entre les ensembles de resumes linguistiques 67

3.5.4 Résumés linguistiques différentiels 70

3.5.5 Travaux de la littératures 72

3.6 Conclusion 74

3.1 Introduction

L'extraction de connaissance à partir de données(ECD) permet de construire des modèles prédictifs à partir de l'observation des données. L'exploitation des données se fait en deux phases : la phase de construction d'un modèle à l'aide d'algorithme d'apprentissage et la phase de déploiement sur de nouvelles données.

Ce chapitre est une introduction aux différentes approches de fouille de données (data mining en anglais) où les différentes étapes du processus d'extraction de connaissances sont décrites à partir des données numériques.

Puis, nous détaillons les différentes méthodes de classification qui traitent les données numériques. Puis, nous présentons les différentes techniques de classification de données (Réseaux de Neurones Artificiels ''RNA '', Support Vector Machine ''SVM'' et K-Plus Proche Voisin ''K-PPV '').
Nous terminons par, un état de l'art sur les différentes modèles qui ont utilisé les résumés linguistiques afin construire un système déduction, ces modèles sont construit à l'aide de calcul de similarité et de distance métrique entre les résumés linguistiques.

3.2 Extraction de connaissance à partir de données

L'extraction de connaissance à partir de données (ECD ou *KDD* pour *Knowledge Discovery in Databases*) permet d'analyser une grande masse de données afin d'extraire une connaissance pertinente .

D'après [74] :
«L'Extraction de Connaissances à partir des Données (ECD) est un processus itératif et interactif d'analyse d'un grand ensemble de données brutes afin d'en extraire des connaissances exploitables par un utilisateur- analyste qui y joue un rôle central»
Où un processus d'ECD est constitué de quatre phases qui sont : *le nettoyage et intégration des données, le prétraitement des données, la fouille de données* et enfin *l'évaluation et la présentation des connaissances.* [75]
Dans ce chapitre, nous nous intéressent sur la phase de la fouille de données car c'est le cœur de ce processus ECD.

3.2.1 Définition de Data Mining (Fouille de données)

La fouille de données est une étape centrale du processus d'extraction de connaissances de base de données (*ECD*). Plusieurs auteurs ont défini le Data Mining, citons les définitions les plus communément admises de Data Mining :
« Le data mining, ou fouille de données, est l'ensemble des méthodes et techniques, destinées à l'exploration et l'analyse de bases de données informatiques (souvent grandes), de façon automatique ou semi-automatique, en vue de détecter dans ces données des règles, des associations, des tendances inconnues ou cachées, des structures particulières restituant l'essentiel de l'information utile tout en réduisant la quantité de données» [76].
Il existe plusieurs tâches qui peuvent être associées au Data Mining où les objets dans ces modèles sont présentés par des enregistrements qui sont constitués d'un ensemble des attributs, ayant des valeurs dans un domaine. Parmi ces taches, citons :

3.2.1.1 L'estimation

Elle consiste à trouver la valeur pertinente d'un attribut pour un but de classification.
Donc, nous attribuons une classe particulière, pour un intervalle de valeurs dans un champ estimé.
Par exemple : d'estimer les revenus des médecins.

3.2.1.2 La prédiction

Cette tâche est proche de l'estimation. Elle consiste à estimer une valeur future. Elle cherche à prédire des valeurs futures d'une action. Par exemple : prédire le départ d'un client à partir aux actions passées.

3.2.1.3 Les règles d'association

Les règles d'association sont liées avec le secteur de la distribution comme par exemple "l'analyse du panier de la ménagère (market basket analysis)», cette méthode de recherche est une association entre les produits .elle permet d'obtenir des informations sur les clients, "qui" sont les clients et "pourquoi" clients font certains achats. Le résultat de cette méthode est un ensemble de règles d'association.

3.2.1.4 La segmentation

La segmentation consiste à former des groupes (clusters) homogènes dans une population (l'ensemble des enregistrements) où il n a pas de classe proposée ou de valeur à prédire (*a priori)*.

3.2.1.5 Classification

La classification est une méthode vaste et utilisable dans plusieurs domaines et services. Car, les méthodes de classification ont pour but, de regrouper les éléments d'un ensemble *X* de nature quelconque, en un nombre restreint de classes.
Cependant, La qualité de la classification peut être jugée sur la base des deux critères :

- Les classes générées doivent être différentes les unes des autres vis-à-vis de certaines caractéristiques,
- Chaque classe doit être plus homogène vis-à-vis ces caractéristiques.

Nous définissons par la suite, les méthodes de classification les plus utilisées.

3.3 La classification

La classification est une discipline scientifique utilisée dans plusieurs domaines de déduction. Les méthodes de classification ont pour but de regrouper les éléments d'un ensemble X dans une classe C. Nous avons un ensemble $X = \{X^1,..., X^n ,..., X^N\}$ et un nombre K optimal de classes, selon leurs ressemblances, où leurs objets appartiennent à des classes spécifiques. [77]
D'une manière générale, les problèmes de classification ont pour tâches à déterminer des procédures permettant d'associer une classe à un objet (individu). Selon Bezdek [78], ces problèmes sont traités par deux variantes ,il y'a la classification dite " *supervisée* " et la classification dite " *non supervisée* ".
Donc, la classification consiste à examiner les caractéristiques d'un objet et lui attribuer une classe. La classe est un champ particulier à valeurs discrètes.

Avant de détailler les méthodes classifications, définissons une phase dite d'apprentissage.

3.3.1 L'apprentissage

Tout d'abord, l'apprentissage est une phase indispensable du modèle de classification supervisée .Il manipule les données et les hypothèses afin de trouver la meilleure hypothèse en fonction des données disponibles.
D'après Vapnik [79] « Le processus d'apprentissage est un processus de choix d'une fonction appropriée à partir d'un ensemble donné de fonctions ».
Alors, l'apprentissage consiste à découvrir des relations pertinentes dans un ensemble de données qui peuvent être utilisées pour le classement des données.
En effet, l'apprentissage peut être supervisé, non supervisé ou semi supervisé :

3.3.1.1 L'apprentissage non supervisé :

L'apprentissage non-supervisé correspond au cas de cible qui n'est pas prédéterminée .L'ensemble de l'entraînement contient seulement les entrées et la nature de fonction f est non définit .Il doit être retournée par l'algorithme d'apprentissage, en parallèle, l'utilisateur doit spécifier le problème à résoudre. Dans tous les cas, le modèle capte certains éléments de la véritable distribution qui génère l'ensemble D (données d'entrées) [80]

$$D = \{(x_i) | \ddot{x}_i \in X\} \text{ Pour } i = 1,\ldots,n \qquad (3.1)$$

Donc, l'apprentissage non supervisé consiste à déterminer les groupes homogènes des exemples car l'utilisateur ne connait pas combien de classe ou de groupe qu'il va obtenu. Ces regroupements se font par le calcul de la similarité.

3.3.1.2 L'apprentissage semi-supervisé (par renforcement) :

L'apprentissage semi-supervisé est quelque sorte d'apprentissage supervisé .Dans cette approche, le réseau doit apprendre la corrélation entrée/sortie à travers une estimation de l'erreur(le rapport échec/succès). Le réseau va donc tendre à maximiser un index de performance qui lui est fourni, appelé signal de renforcement. Donc, le système peut savoir si la réponse qu'il fournit est correcte ou non.[81]

3.3.1.3 L'apprentissage supervisé :

L'apprentissage supervisé consiste à trouver des caractéristiques générales d'un ensemble d'objets .Il traite les données de la forme (x, y) où x représente un objet et y la valeur associée à cet exemple (continue ou discrète) c.à.d. y est une variable discrète finie qui indique l'étiquette associée à l'exemple x.
L'objectif de l'apprentissage supervisé est d'extraire une relation reliant y à x. Cette relation est représentée par la fonction f. $f : \mathrm{Y} \rightarrow X$
En utilisant le principe d'induction, nous extrayons la fonction f à partir de l'ensemble d'apprentissage D qui est définit comme suit :

$$D = \{(\mathrm{x}_i, \mathrm{y}_i) \mid \mathrm{x}_i \in \mathrm{X},\ \mathrm{y}_i \in \mathrm{Y}\} \quad \text{Pour i=1...N} \qquad (3.2)$$

D correspond à l'ensemble de N paires d'entrées x_i et de cibles associées à y_i .
Donc, l'apprentissage supervisé est basé sur l'induction qui est, en réalité, le passage d'un cas particulier à une généralisation. Pour déterminer la classe d'un exemple, les algorithmes inductifs introduisent des contraintes dans d'apprentissage. Ils peuvent faire des suppositions sur la définition de l'hypothèse reliant les objets et leur classes[82][83].[84]

Nous concluons que la phase d'apprentissage supervisé , consiste à analyser les ressemblances entre les formes d'une même classe et les dissemblances entre les formes de classes différentes , pour en déduire les meilleures séparations possibles entre les classes.

3.3.2 Principe de la classification

La classification est liée à la notion de partition d'un ensemble fini.
Plusieurs auteurs ont définis la notion de classification, nous avons retenu les plus intéressantes :

- *« La classification consiste à attribuer des objets, des candidats, des actions potentielles à des*

catégories prédéfinies » Henriet [85].

- *« La classification est l'action de regrouper en différentes catégories des objets ayant certains points communs ou faisant partie d'un même concept sans avoir connaissance de la forme ni de la nature des classes au préalable. On parle alors de problème d'apprentissage non supervisé ou de classification automatique. L'action d'affecter des objets à des classes prédéfinies on parle dans ce cas de classification supervisée ou de problèmes d'affectations »* [86]

3.3.2.1 Définition d'une classe

La classe est une instance déterminée par l'attribut «classe» où ensemble des instances d'apprentissage sont utilisées dans la construction du modèle.
Nous avons un couple (x, y), où $x \in X$ est la description ou la représentation de l'objet et $y \in Y$ représente la classe attribuée à x.
La classe y appartient à l'ensemble $Y = \{y_1, \ldots, y_C\}$, C désigne le nombre de classes auxquelles l'objet x appartient.

3.3.2.2 Définition d'un classifieur

On appelle un classifieur f : tous les applications mesurables où $f : X \rightarrow Y$, la qualité de f est évaluée par sa capacité de généralisation. Nous estimons le risque réel noter par $R(f)$ qui est défini par :

$$R(f) = \int_{X \times Y} L(f(x), y) dF(x, y) \qquad (3.3)$$

Où $L(f(x), y)$ est la fonction de perte qui évalue le coût de la décision de $f(x)$. $F(x, y)$ est la fonction de répartition des données.
Dans tous les cas, on a :

$$R(f) = P(y \neq f(x)) \qquad (3.4)$$

Le risque d'une fonction f est un calcul de probabilité où le classifieur f prédise une réponse différente de celle du superviseur.
Le problème général de la classification supervisée peut s'exprimer de la manière suivante : un échantillon $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$,il faut trouver une fonction f qui minimise le risque $R(f)$[84] .

Il existe trois types de classification (classification supervisé, classification non supervisé et la classification semi_supervisé).Dans notre thèse, nous intéressons surtout au système de classification supervisé.

3.3.3 Classification supervisée

La classification supervisée est très utilisée dans le domaine médical, elle détermine l'état du patient en connaissant l'état des autres. Nous évaluons les risques des maladies pour les nouveaux patients. . [87].

La classification est de crées des modèles prédictifs, à partir de données étiquetés grâce à un algorithme d'apprentissage.

Nous appelons $X = \{X_1, X_2, ..., X_d\}$ le vecteur de **d** variables explicatives. La variable cible (classe à prédire) est : $C = \{c_1, c_2, ..., c_k\}$.

Nous avons $x : x_1, x_2, ...x_d, c_k$ où x_i correspondent aux **d** variables descriptives et c_k est une classe (étiquette) à prédire.

Les variables explicatives peuvent être numérique : $x_i \in R$, ou catégorielles $x_i \in \{\upsilon_1, \upsilon_2, ..., \upsilon_v\}$. Un classifieur est un modèle **f** qui détermine la classe d'un exemple : $f : X \to C$ [88]

Alors, pour construire un modèle de classification supervisé, il faut passer par trois étapes importantes :

- Construction d'un modèle à partir de l'ensemble d'apprentissage (training set).
- L'évaluation de la qualité/précision du classifieur.
- L'utilisation du modèle pour la classification de nouvelles instances.

Nous détaillons par le schéma (figure 3.1) qu'il explique ces étapes.

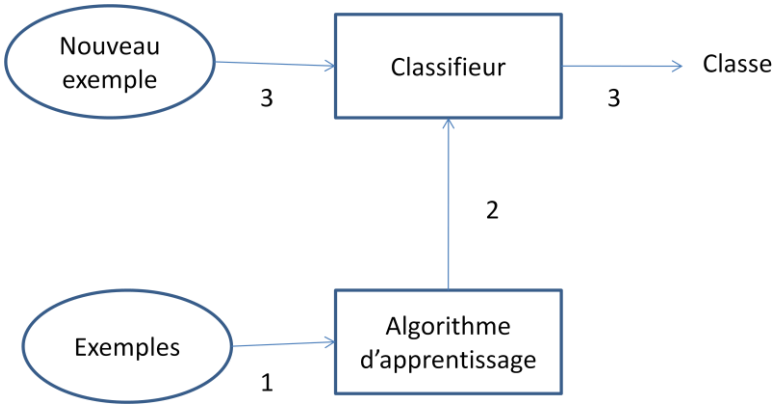


Figure 3-1Les étapes de système de classification

Ce schéma présente les principales étapes du système de classification où la première étape est d'établir l'un des d'algorithmes d'apprentissage afin de construire une base de connaissance .Cette base de connaissance est essentiel, pour prédire un nouveau exemple à l'aide d'un classifieur.

Alors, l'objectif de la classification est de fournir une procédure ayant un pouvoir prédictif et garantissant des prédictions fiables sur les nouveaux exemples.

La qualité prédictive d'un modèle peut être évaluée, en calculant le risque empirique qui permet de mesurer la probabilité d'une mauvaise classification d'une hypothèse donnée.il existe pour cela plusieurs calculs.

3.3.3.1 Risque réel

Soit f est une hypothèse prise à partir d'un échantillon S d'objets. Le risque réel de $\boldsymbol{f}$ est défini par l'équation :

$$R(f) = \int_{X \times Y} L(f(x), y) dF(x, y) \qquad (3.5)$$

Où L est une fonction de perte ou le coût associé à la mauvaise classification. L'intégrale prend en compte la distribution F de l'ensemble des objets sur le produit cartésien $X \times Y$. La fonction de perte la plus simple utilisée en classification est définie par :

$$L(f(x), y) = \begin{cases} 0 si \ y = f(x) \\ 1 si \ y \neq f(\mathrm{x}) \end{cases} \qquad (3.6)$$

La distribution des exemples étant inconnue, ce qui rend impossible le calcul du risque réel. Le système d'apprentissage n'a en effet accès qu'à l'erreur apparente (erreur empirique) qui est mesurée sur l'échantillon d'apprentissage [89]

3.3.3.2 Risque empirique

Soit S l'ensemble d'apprentissage $S = \{(\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_n, y_n)\}$.

f Une hypothèse, le risque empirique de f est définie par l'équation :

$$R_{emp}(f) = \frac{1}{N} \sum_{i=1}^{N} L(f(x_i), y_i) \qquad (3.7)$$

Le risque empirique est le nombre moyen des objets mal classés.
Lorsque la taille d'un l'échantillon tend vers l'infini, le risque apparent converge vers le risque réel si les éléments de S sont tirés aléatoirement, à partir de l'ensemble d'apprentissage.[89].

Nous citons dans la section suivante l'état de l'art sur les différentes méthodes de classification les plus utilisées.

3.4 Les méthodes de classification

3.4.1 Séparateurs à Vaste Marge

SVM (Support Vector Machine ou Séparateurs à Vaste Marge) est l'une des techniques de la classification. Le principe est minimiser l'erreur empirique de classification et de maximiser la marge géométrique entre les classes, «Séparateur à Vaste Marge».

Les bases des Séparateurs à Vaste Marge (SVM) ont été proposées par Vapnik en 1963 où Vapnik est le fondateur de l'idée de base de cette méthode. Le principe est de trouver l'hyper plan qui maximise la distance entre les objets de différentes classes.
. Depuis les années 1995, les chercheur ont amélioré l'étude des méthodes à base de SVM [90][91][92], dans le domaine pratique et théorique[79][93][94][95].

Cette méthode a montrée ses preuves dans plusieurs domaines comme la reconnaissance de chiffres manuscrits, la classification de textes ou la bioinformatique et sur des ensembles de données de très grandes dimensions.

3.4.1.1 Principe

Les entrées X sont transformées en un vecteur dans un espace de Hyper plan F . Dans le cas d'un classement en 2 classes, nous déterminons un hyperplan dans cet espace F . La solution optimale repose sur la propriété que les objets sont les plus éloignés possibles de l'hyperplan, on maximise ainsi les marges.

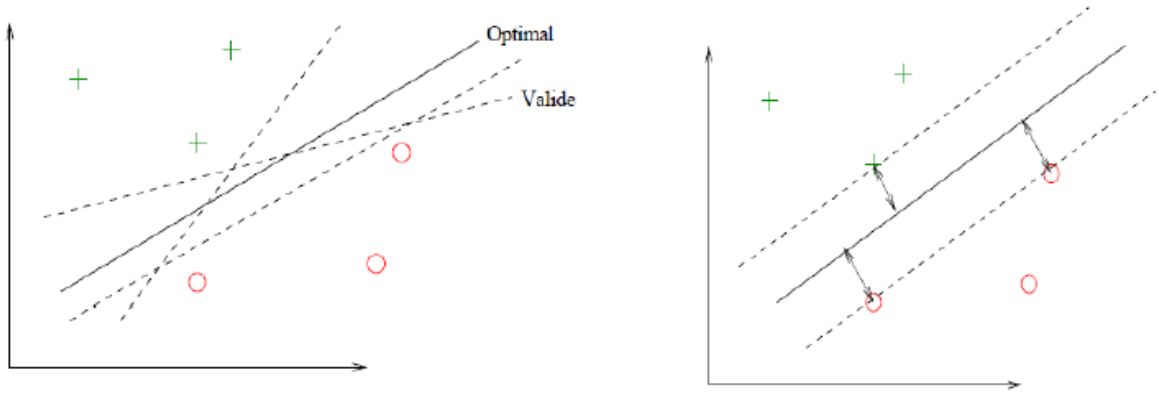


Figure 3-2Hyperplan séparateur

Soit x jeu d'entrées, produit une sortie : y = f(x).
Le but est de trouver f à partir de l'observation d'un certain nombre de couples (entrée, sortie).
Formellement, nous allons supposer des couples (x_i, y_i) où $1 \le i \le N$, tel que $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ est l'ensemble de données linéairement séparables (Figure 3.2) avec
$x \in \mathbb{R}^n$ et $y \in \{-1,1\}$.L'appartenance d'une observation x_i a une classe ou une autre. Elle est notée par la valeur -1 ou 1 dans son étiquette y_i .
La classification consiste à déterminer un hyperplan qui sépare deux ensembles de points, cet hyperplan est définie par :

$$\forall (x_i, y_i) \in D : f(x_i) = \langle w, x_i \rangle + b = 0 \qquad (3.8)$$

Il est caractérisé par deux grandeurs :

- vecteur normal w (vecteur de poids) .
- biais b qui permet d'engendrer une simple translation de l'hyperplan de séparateur, la valeur $\frac{b}{\|w\|}$ représente la distance perpendiculaire de l'hyperplan à l'origine.

Toute observation x_i est affectée à la classe qui correspond au signe de

$$f(x_i) \ : \langle w, x_i \rangle \ge +1 \text{ pour } y_i = +1$$
$$\langle w, x_i \rangle \le -1 \text{ pour } y_i = -1 \qquad (3.9) \qquad [84]$$

3.4.1.2 Ajustement

Pour maximiser la marge il faut minimiser ||w || ou ||w ||²/2 sous les contraintes de y_i *où* $f(x_i) \geq 1$.

Nous utilisons la méthode des multiplicateurs de Lagrange en prenant en compte que les vecteurs ***xi*** les plus proches de l'hyperplan (vecteurs supports).

Lorsque tous les cas ne sont pas séparables, nous introduisons un terme d'erreur : y_i *où* $f(x_i) \geq 1-\xi_i$ .La transformation en vecteur ne fait intervenir que le produit scalaire dans **F**.

Nous recherchons directement l'expression du produit scalaire à partir des coordonnées initiales à l'aide d'une fonction *k* appelée noyau. [96]

3.4.1.3 Avantages et inconvénients

Les avantages :

- l'apprentissage peut être indépendant de la dimension du vecteur d'entrées.
- la solution obtenue à partir de l'algorithme SVM correspond à la recherche d'une hypothèse, possédant de bonnes capacités de généralisation, à partir d'un espace d'hypothèses donné.
- l'algorithme d'apprentissage reste identique.

Les inconvénients :

- les résultats de SVM ne sont pas facilement interprétables. Les seules informations fournies sont des vecteurs supports sans aucune autre indication.
- la fonction noyau est souvent des paramètres libres.
- L'hyperplan sélectionné, même géométriquement, la marge est maximum et faible.
- la sélection d'une mauvaise fonction de paramètres peut produire un sur-apprentissage.

3.4.2 Arbre de décision

La méthode des arbres de décision est l'une des méthodes la plus utilisé de data mining, elle fournit des règles explicites de classement et supporte bien les données hétérogènes, manquantes et non linéaires. Cette méthode est préférée dans la prédiction du risque par ce qu' elle est robuste, elle est donc à la catégorie des classifications hiérarchiques descendantes supervisées.[97]

Les arbres de décision correspondent à un ensemble d'algorithmes (CART [98],ID3 [99], C4.5 [100], CHAID [101], etc.) et sont très utilisés depuis de nombreuses années dans le cadre de l'apprentissage supervisé [102]. Les chercheurs ont proposées des versions évoluées d'algorithmes à base d'arbres de décision exploitant ,le principe de minimisation du risque structurel et la notion de VC dimension [103] afin d'améliorer les capacités de généralisation des arbres de décision produits.[104].

Alors, les arbres de décision peuvent traiter des données représentées par des attributs quantitatifs, des attributs qualitatifs ou des représentations composites.

3.4.2.1 Principale de l'arbre de décision

La technique de l'arbre de décision est d'employer un classement pour détecter des critères qui permettent de répartir les individus d'une population en n classes qui est souvent n=2. Nous commençons par choisir la variable qui sépare les individus de chaque classe, de façon à avoir des sous-populations qu'on appelle nœuds où une seule classe peut contenir, chacune, le plus possible d'individus. Puis, on refait la même opération sur chaque nouveau nœud jusqu'à ce que la séparation des individus ne soit plus possible.

Les nœuds terminaux (les feuilles) sont tous constitués d'individus d'une seule classe qui vont satisfaire l'ensemble des règles. L'ensemble de ces règles constitue le modèle de classement (Figure 3.3.)[97]

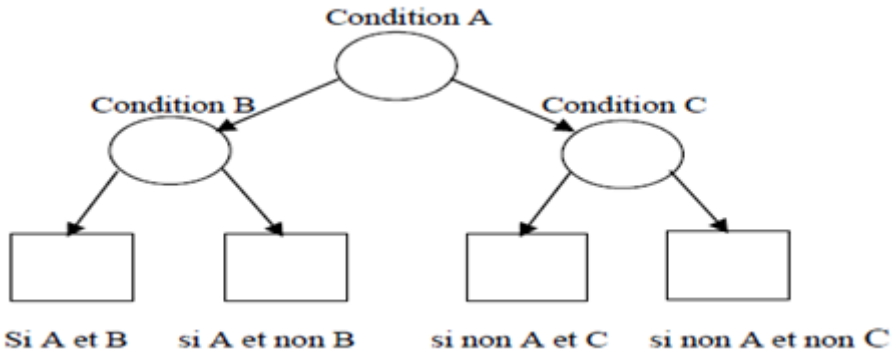


Figure 3-3Arbre de décision

Pour construire l'arbre de décision, il faut définir une suite de nœud où chaque nœud permet de faire une partition des objets en 2 groupes, sur la base d'une des variables explicatives. Il convient de:

- Définir un critère permettant de **sélectionner le meilleur nœud** possible.
- Définir l'arrêt de découpage qui permet de définir le **nœud terminal** (feuille),
- Attribuer au nœud terminal la classe ou la valeur la plus probable.
- Sélectionner un sous arbre optimal quand le nombre de nœud devient très important (l'arbre maximal).
- Valider l'arbre à partir d'une technique de validation croisée ou d'autres.

3.4.2.2 Avantages et inconvénients

Les avantages :

- Adaptabilité aux attributs de valeurs manquantes
- Bonne lisibilité du résultat.
- Traitement de tout type de données.
- Sélection des variables pertinentes
- Classification efficace(les algorithmes de génération d'arbres de décision sont disponibles dans tous les environnements de fouille de données).
- Résolution des tâches d'estimation et de prédiction.

Les Inconvénients :

- Sensible au nombre de classes.
- Moins performant (très complexes lorsque le nombre de variables mises en jeu et les classes augment).
- Peut lisible et un mauvais classement dans le cas d'un arbre trop détaillé.
- Pas de traitement de flou de données (méthode non incrémental)

3.4.3 Réseaux de neurone

Les réseaux de neurones artificiels (Artificial Neural Networks) sont des modèles schématiquement inspirée du fonctionnement des neurones biologiques [105].

Actuellement les méthodes connexionnistes sont très nombreuses et elles partagent des points communs avec les principes d'autres types d'algorithmes d'apprentissage [106][107].

«Les réseaux de neurones sont des outils très utilisés pour la classification, l'estimation, la prédiction et la segmentation. Ils sont issus de modèles biologiques, constitués d'unités élémentaires (les neurones) et organisées selon une architecture»[108] .

3.4.3.1 Principe

Cette méthode repose sur la notion de neurone formel. Un neurone formel est un modèle caractérisé par des signaux d'entrée (variables explicatives),et par une fonction d'activation $f, \mathrm{f}(\alpha_0 + \sum_i \alpha_i \times x_i)$.*f* peut être linéaire, à seuil stochastique et plus souvent sigmoïde.

Le calcul des paramètres se fait par apprentissage.

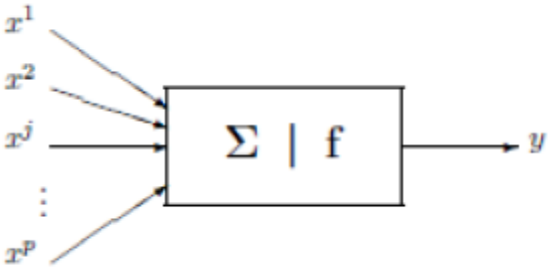


Figure 3-4Forme général de réseau de neurone

Les neurones sont associés en couche :

- Une couche d'entrée lit les signaux entrée, noter par x^j .
- une couche de sortie qui fournit la réponse du système.
- Une ou plusieurs couches cachées participent au transfert. Un neurone d'une couche cachée est connecté ,en entrée, à chacun des neurones de la couche précédente ;et en sortie à chaque neurone de la couche suivante.

..[96]

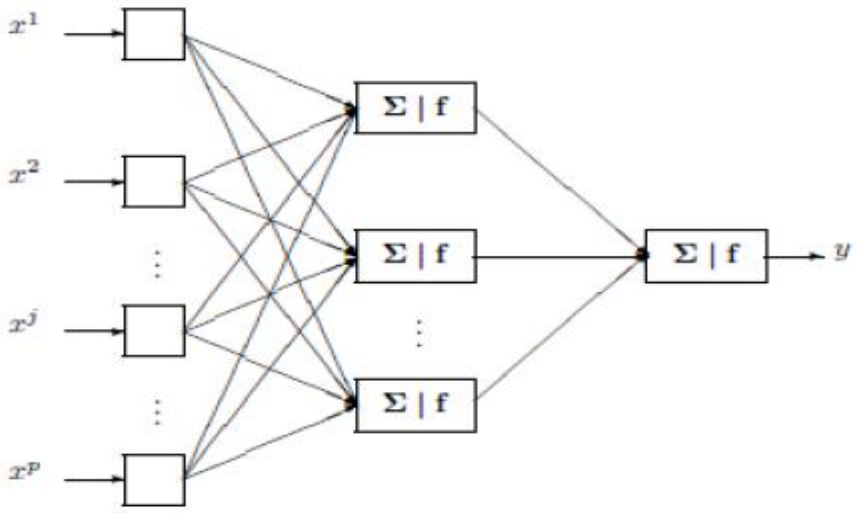


Figure 3-5 Multi couche de réseau neurone

Lorsque le réseau est utilisé dans une technique de prédictive, il y a une ou plusieurs variables d'entrées correspondant, chacune à un nœud entré et un *nœud de sorti.*
La couche d'entrée et la couche sortie sont connectés à des nœuds appartenant à un niveau intermédiaire, appelée *la couche cachée.* Il peut exister plusieurs couches cachées [109].
La dernière couche est constituée d'un seul neurone, muni de la fonction d'activation identité. Tandis que les autres neurones (couche cachée) sont munis par la fonction sigmoïde.

3.4.3.2 Mise en œuvre

Les étapes dans la mise en œuvre d'un réseau de neurones, pour la prédiction ou le classement sont :

- Identification des données en entrée et en sortie.
- Normalisation de ces données.
- Constitution d'un réseau avec une structure adaptée.
- Apprentissage du réseau.
- Test du réseau.
- Application du modèle génèreront l'apprentissage.
- Dé normalisation des données en sortie.

3.4.3.3 Avantages et Inconvénients

Les avantages :

- Lisibilité du résultat(les réseaux traitent facilement les données réelles)
- L'algorithme est robuste au bruit.
- L'efficacité et la rapidité dans le modèle de Classification
- Combiner avec d'autres méthodes de classification.

Les inconvénients *:*

- L'échantillon nécessaire à l'apprentissage doit être suffisamment grand et représentatif des sorties attendues. Il faut passer un grand nombre de fois tous les exemples de l'échantillon d'apprentissage avant de converger et donc le temps d'apprentissage peut être long.
- L'apprentissage n'est pas incrémental
- Le codage des entrées (toutes les entrées doivent se trouver dans un intervalle défini entre 0 et 1. Ce qui amène à des transformations et risquent de fausser les résultats.).
- Le choix des valeurs initiales à des poids synaptiques du réseau.
- Réglage du pas d'apprentissage (choisi soigneusement pour s'assurer des bons résultats)**.**

3.4.4 Les plus proches voisins

La méthode ''des plus proches voisins'' (*PPV* en bref, *nearest neighbor* en anglais ou *NN)est* la méthode de classification la plus utilisée. Elle a un échantillon d'apprentissage à base d'instances.
La construction de ce modèle se fait à partir d'un échantillon d'apprentissage qui est associé à une fonction de distance et à une fonction de choix de la classe, en fonction des classes des voisins les plus proches. Donc, l'idée de base de cette méthode est de prendre la décision à partir des cas similaires.

L'algorithme de classification de la méthode *PPV* est :[97]

Algorithme 3.1 de classification K-PPV
Paramètre :nombre K VOISIN
Donné :échantillon de m exemples et classes
Classe d'un exemple X est c(X)
Entrée :un enregistrement Y
Déterminer K-plus proche de Y en calculant la distances
Combiner les classes de K exemple dans classe c
Sortie :la classe de Y est c(Y)=c

Selon l'algorithme *k*-PPV, il y'a deux aspects importants :

- d'une part, il est nécessaire de parcourir l'ensemble de la base d'apprentissage. .
- d'autre part, choisir une fonction de distance utilisée par l'algorithme pour mesurer la proximité des objets c.à.d., le bon choix de la fonction de distance/similarité pour nous donner un meilleur taux de classification.

3.4.4.1 Les distances

La distance est une application qui permet de regrouper un groupe d'individus similaire et de séparer les individus qui ne présentent aucune similarité.
La fonction de distance *d* entre deux points *xi* et *xj* est une application de $\mathbb{R}d \times \mathbb{R}d$ dans $\mathbb{R}+$ satisfaisant les propriétés suivantes :

- Non négativité : $d(x_i, y_j) \geq 0$
- Symétrie : $d(x_i, y_j) = d(x_j, y_i)$
- Séparation : $d(x_i, y_j) = 0 \Rightarrow x_i = x_j$
- Minimalité : $d(x_i, y_j) = 0$.
- Intégralité triangulaire : $d(x_i, y_j) \leq d(x_i, y_k) + d(x_k, y_j)$

Plusieurs fonctions de distance ont été définies dans la littérature. Ces fonctions ont une valeur proche de 0 pour un couple de points rapprochés , et proche de+∞ pour un couple de points très éloignes.[84]
Les distances les plus utilisées sont :

La distance de Manhattan

La distance de Manhattan (City block distance) est associée à la norme 1. Pour deux vecteurs x_i^t et x_j^t , la distance de Manhattan est définie comme suit :

$$d(x_i^t \times x_j^t) = \sum_{t=1}^{n} \left| x_i^t - x_j^t \right| \qquad (3.9)$$

La distance de Manhattan est la somme des différences absolues. [84]

La distance Euclidienne

La distance euclidienne est la plus utilisée, elle est définie par :

$$d(x_i^t \times x_j^t) = \sqrt{\sum_{t=1}^{n} \left(x_i^t - x_j^t \right)^2} \qquad (3.10)$$

Notons que la distance Euclidienne est un cas particulier de la distance de Minkowski avec *p=2*. [84]

La distance Minkowski (p-distance)

La distance de Minkowski est une généralisation de la distance Euclidienne et la distance de Manhattan. Elle est définie comme suit :

$$d(x_i^t \times x_j^t) = \sqrt[p]{\sum_{t=1}^{n} \left| x_i^t - x_j^t \right|^p} \qquad (3.11)$$

Avec p = 1,2, … , ∞.

La distance Tchebychev (∞ –distance)

La distance de Tchebychev est le maximum des valeurs absolues :

$$d(x_i^t \times x_j^t) = \lim_{p\to\infty} \sqrt[p]{\sum_{t=1}^{n} \left| x_i^t - x_j^t \right|^p} = \max_{1\le t\le n} \left| x_i^t - x_j^t \right| \qquad (3.12)$$

La distance Cosinus

La distance cosinus est décrite comme suit :

$$d(x_i^t \times x_j^t) = \frac{\sum_{t=1}^{n} x_i^t \times x_j^t}{\sqrt{\sum_{t=1}^{n} \left(x_i^t\right)^2} \times \sqrt{\sum_{t=1}^{n} \left(x_j^t\right)^2}} \qquad (3.13)$$

La corrélation

La fonction de corrélation de deux variables x_i^t et x_j^t est donnée par :

$$d(x_i^t, x_j^t) = 1 - C_{x_i^t, x_j^t} \qquad (3.14)$$

Où *c* est le coefficient de corrélation de Pearson. [84]

Les distances : Euclidienne, Manhattan, Cosinus et de Corrélation sont les types de distances les plus utilisées dans l'algorithme des *k* plus proches voisins.

3.4.4.2 Avantages et inconvénients

Les avantages :

- Absence de phase d'apprentissage.
- Résultats trouvés claires.
- traitements des données hétérogènes avec un grand nombre d'attributs.

Les inconvénients :

- Difficultés de sélection des attributs pertinents.
- Lenteur du temps de classification.
- Dépendance des performances de la méthode en fonction du choix de la distance, du nombre de voisins et du mode de combinaison des réponses des voisins.

Enfin, Il faut noter qu'il y'a aussi des algorithmes de classification qui utilisent la notion de la logique floue comme :l'algorithme des C-moyennes floues (J C Bezdek, 1981, [111][112]))et qui sont basés sur le système d'inférence floues. Mais, le problème majors de ces SIF (Systèmes Inférence flou) est au niveau de l'augmentation le volume de la base de données et qui provoque le problème de l'augmentation les règles floues du système.

Dans la section suivante, nous parlerons sur la relation entre le système prédiction et les résumés linguistiques où ces résumés utilisent la notion de la logique floue.

3.5 Prédiction et le calcul de la distance métrique des résumés linguistique

Le but de la création des résumés linguistiques est de condenser sémantiquement les informations. La plupart des études de développement et d'amélioration de ces méthodes de résumés sont consacrées pour la création des systèmes interrogations flexibles.

Le résumé de Yager est la méthode la plus exploitée, mais les informations produites par ce modèle sont volumineuses, car ils contiennent des redondances (plusieurs résumés ayant même poids), et des informations sans poids.

C'est pour cette raison que nous trouvons dans la littérature l'utilisation de calcul de la similarité entre deux résumés linguistiques, afin de comparer leur ressemblance. Si deux résumés sont similaires, alors ils ont le même poids d'informations. Ceci nous amène à garder l'un des deux.

Nous détaillerons, par la suite, les différentes présentations de protoforme de résumé linguistique qui ont été proposé. Nous définissions le calcul de similarité entre deux résumés linguistique. Enfin, nous présenterons un état de l'art sur les modèles qui ont utilisé ces calculs.

3.5.1 Présentation de résumé linguistique

3.5.1.1 Les protoformes classiques :

Les résumés linguistiques des données sont généralement de courtes phrases en langage naturel et qui sont développés à partir de données numérique. Ces données brutes sont volumineuses et incompréhensibles pour un humain.

La notation la plus utilisée pour présenter une protoforme de résumé linguistique est présenté par [3]:

$Y = \{y_1, y_2, \ldots, y_n\}$ est l'ensemble des objets (enregistrements). Par exemple : les patients dans une collection.

$A = \{A_1, A_2, \ldots, A_m\}$ est l'ensemble des attributs caractérisant les objets de Y. exemple (Age, tension….).

Fondamentalement, un résumé linguistique est linguistiquement quantifié, il peut être écrit sous deux formes différentes :

- une simple protoforme :

Qy's are P (3.15)

- l'extension de protoforme est:

QRy's are P (3.16)

Un résumé linguistique contient:

- Un résumé **P** : valeur linguistique d'un attribut (prédicat flou) définie sur le domaine de l'attribut A_j (par exemple, jeune pour l'attribut l'Age).
- Une quantité **Q** : quantificateur linguistique (par exemple, la plupart);
- Une vérité (validité) **T** du résumé : nombre évaluant le degré vérité (validité) du résumé, et appartenant à l'intervalle [0,1] (par exemple, 0,7);
- Un qualificateur **R** : deuxième variable linguistique (prédicat flou) d'un autre attribut, définie sur le domaine de l'attribut A_k .

Q, P et **R** sont modélisés par des ensembles flous sur des domaines appropriés. La vérité (validité) **T** d'un résumé linguistique correspond directement à la valeur de vérité dans les équations (3.15) et (3.16).

Le degré de vérité peut être calculé en utilisant, soit le calcul de propositions quantifiées de Zadeh [113], soit d'autres interprétations de quantificateurs linguistiques cité ci-dessous.

Les valeurs de vérité des équations (3.15) et (3.16) sont calculées, respectivement, comme suit :

$$T(\textbf{Qy's are P}) = \mu_Q\left(\frac{1}{n}\sum_{i=1}^{n}\mu_p\left(y_i\right)\right) \qquad (3.17)$$

$$T(\textbf{QRy's are P}) = \mu_Q\left(\frac{\sum_{i=1}^{n}\mu_P(y_i)\wedge\mu_R(y_i)}{\sum_{i=1}^{n}\mu_R(y_i)}\right) \qquad (3.18)$$

Où ∧ est le fonctionnement minimum (il peut être un autre opérateur approprié noter t_norme), et Q est un ensemble flou qui représente le quantificateur linguistique qui est normal et monotone. [114]

N.B : Pour combiner plus d'une valeur d'attribut, comme P1 et P2, il faut utiliser le t_norme de conjonction (par exemple, le minimum ou le produit) ou le s-norme de la disjonction de correspondante (par exemple : maximale ou la somme probabiliste).

En effet, il n'y a pas que le degré de vérité pour mesurer la qualité de résumé linguistique mais aussi autres critères de qualité sont utilisés, comme le degré de focalisation (focus), introduit par Kacprzyk et Wilbik [115].

Le but de degré de focalisation est de chercher les meilleurs résumés linguistiques, en prenant en compte les informations supplémentaires avec les valeurs de vérité correspondantes. Le calcul de degré de focalisation de cette forme existe uniquement pour les résumés de protoforme étendue (type 2) car le degré de focalisation donne une satisfaction approprié de R pour tous les objets.

Le rôle du degré de focalisation est similaire au concept de support dans les règles d'association qui sont utilisé dans un contexte de data mining. Il fournit une mesure que nous considérons comme valeur de vérité fondamentale. Il peut nous aider également à contrôler le processus de rejet des résumés linguistiques non prometteurs.

$$d_{foc}(\mathbf{QRy's\ are\ P}) = \frac{1}{n}\sum_{i=1}^{n}\mu_R(y_i) \qquad (3.19)$$

Si le degré de focalisation est élevé, alors le résumé contient de nombreux objets (enregistrements), c.à.d. il est plus général.

Ces deux critères (degré de vérité et degré de focalisation) ne sont pas les seuls à valider un résumé. Dans [116], il y a une autre approche pour générer ces résumés linguistiques et qui fourni les règles sur la façon de rejeter les résumés non intéressants ou non prometteurs.

3.5.1.2 L'extension de protoforme de Yager

3.5.1.2.1 Protoforme de contextualités temporelle

Les protoformes classiques proposées par Yager ne contiennent pas la nature temporelle des données .citons le cas d'une base composée d'observations sur de multiples phénomènes qui sont observés sur de longues périodes pour les mêmes objets étudiés. Exemple : «La plupart des patients ont une pression artérielle élevée dans la plupart du temps».

Chaque objet y_i consiste à une série chronologique, qui peut s'écrire :

$y_i = (y_{it})_{t=1..\tau}$ où i = 1…n.

Cette protoforme de résumé contient une contextualité temporelle des objets avec des différents attributs.

La protoforme ,proposé par Almeida [117] , a été étendu à partir des protoformes originaux de Yager:

$$Qy's\ are\ PQ_\tau times \qquad (3.20)$$

$$QRy's\ are\ PQ_\tau times \qquad (3.21)$$

où Q_τ est une quantification temporelle.

Évaluation proposée du degré de vérité

Afin d'évaluer la validité des résumés temporels de l'équation (3.20), le degré de vérité est :

$$T = \mu_Q\left(\frac{1}{n}\sum_{i=1}^{n}\mu_{Q_\tau}\left(\frac{1}{\tau}\mu_P(y_{it})\right)\right) \qquad (3.22)$$

Cela signifie que la valeur $\mu_P(y_i)$ a été remplacée par l'évaluation de $\mu_{Q\tau}$ .L' équation (20) utilise la moyenne μ_P sur une série temporelle.

Par l'exemple "**Peu de patients ont une fréquence cardiaque basse dans la plupart du temps**". Pour chaque patient, il faut donner d'abord la fuzzification d'attribut ''rythme cardiaque'', puis la quantité temporelle pour la fréquence cardiaque basse. Puis, nous quantifions le nombre de patients ayant une fréquence cardiaque basse dans la plupart du temps.

Pour la protoforme étendue du type 2, le calcul de la valeur de vérité s'écrit :

$$T = \mu_Q\left(\frac{1}{\tau}\sum_{i=1}^{n}\mu_{Q\tau}\left(\frac{\sum_{t=1}^{\tau}\mu_P(y_{it})\wedge\mu_R(y_{it})}{\sum_{t=1}^{\tau}\mu_R(y_{it})}\right)\right) \qquad (3.23)$$

3.5.1.2.2 Contextualisation par catégorie

Dans ce travail, Almeida a proposé également une extension catégorique des résumés linguistiques: il a utilisé des étiquettes de catégories nettes $C = \{c_1, \ldots, c_K\}$ .comme une formule qui donne un aperçu de la différence qu'il y a entre les patients et les événements dans les données médicales.

Pour cela, il a proposé d'étendre les protoformes simples et complexes sous la forme de:

$$Q\ y\text{'s with c are P} \qquad (3.24)$$

$$Q\ Ry\text{'s with c are P} \qquad (3.25)$$

Exemple «La plupart des patients atteints de la maladie X ont une pression artérielle basse».

Les protoformes (3.24) et (3.25) peuvent être étendues pour fournir une contextualisation temporelle et catégorielle:

$$Q\ y\text{'s with c are P } Q_\tau\ time \qquad (3.26)$$

$$Q\ Ry\text{'s with c are P } Q_\tau\ time \qquad (3.27)$$

Exemple "La plupart des patients atteints de la maladie X ont une fréquence cardiaque basse dans la plupart du temps".

Nous définissons pour cela le label de catégorie c dans le sous-ensemble de Y comme :

$$Y^c = \{y_i \in Y / y_i \in c\} \qquad (3.28)$$

Donc, $Y = Y^{c1} \cup Y^{c2} \cup \ldots \cup Y^{ck}$.

Alors, La valeur de vérité pour (3.24), (3.25), (3.26) et (3.27) peut être obtenue en substituant Y par Y^c et n par n^c ,tel que :

$$T = \mu_Q\left(\frac{1}{n^c}\sum_{y_i \in Y^c}\mu_P(y_i)\right) \qquad (3.29)$$

Où objets yi appartiennent à une classe donnée c (catégorie c). n^c est le nombre d'éléments de cet ensemble.

3.5.2 Evaluation de la similarité d'une protoforme classique :

Nous détaillons, dans ce sous-titre, le calcul le degré de similarité qui est utilisé entre deux résumés linguistiques. Nous avons deux protoformes de résumés linguistique, proposé par Yager :

Q1R1 y sont P1

Q2R2 y sont P2

Avec **T1** et **T2** les valeurs de vérités, et R1 et R2 les qualificatifs, peuvent être vides dans le cas d'une simple protoforme .

Le degré de similarité est calculé par le minimum entre les quatre éléments suivant:
$sim\ (P_1,\ P_2), sim\ (Q_1,\ Q_2), sim\ (T_1,\ T_2), sim\ (R_1,\ R_2)$

- La similarité de **P1** et **P2 :** La formule de calcul de cette similarité dépend si l'utilisateur décrit les mêmes attributs ou non :

$$sim(P_1, P_2) = \min\left(\frac{a}{b}, \frac{\int\left(\mu_{P_1} \cap \mu_{P_2}\right)}{\int\left(\mu_{P_1} \cup \mu_{P_2}\right)}\right) \qquad (3.30)$$

où **a** est le nombre des attributs communs pour les résumés P_1 et P_2 et **b** est le nombre des attributs distincts utilisés dans le résumé P_1 ou P_2.

Le rapport $\frac{a}{b}$ est une mesure de Jaccard [118] des ensembles d'attributs pour les résumés P1 et P2.

$\frac{\int\left(\mu_{P_1} \cap \mu_{P_2}\right)}{\int\left(\mu_{P_1} \cup \mu_{P_2}\right)}$ est également la mesure de Jaccard des résumés où μ_{P_1} et μ_{P_2} sont les fonctions d'appartenance des ensembles flous utilisés pour modéliser les résumés P_1 et P_2**.**

Sachant que, la distance de Jaccard est définie par :

$$d\left(A,\ B\right) = \frac{1 - \left|A \cap B\right|}{|\ A \cup B\ |} \qquad (3.31)$$

Si $\left|A\right|$ indique la cardinalité de l'ensemble A (cardinalité standard pour les ensembles, *Sigma-count* pour les ensembles flous) Alors , on peut considérer la distance d'un résumé linguistique comme métrique.

- La similarité des quantificateurs Q_1 et Q_2 est calculé en utilisant la mesure de

$$\text{Jaccard} : sim(Q_1, Q_2) = \frac{\int\left(\mu_{Q_1} \cap \mu_{Q_2}\right)}{\int\left(\mu_{Q_1} \cup \mu_{Q_2}\right)} \qquad (3.32)$$

où μ_{Q_1} et μ_{Q_2} sont les fonctions d'appartenance des ensembles flous, utilisés pour modéliser les quantificateurs linguistiques Q_1 et Q_2.

• La similarité des valeurs de vérité T_1 et T_2 est calculée comme :

$$sim(\mathrm{T}_1, T_2) = 1 - \left|\mathrm{T}_1 - T_2\right| \qquad (3.33)$$

Où $\left|\mathrm{T}_1 - T_2\right|$ est la valeur absolue de la différence.

• Similarité des qualificateurs R_1 et R_2 . R doit être considéré comme étant un ensemble flou qui caractérise l'univers entier de Y, c'est-à-dire l'ensemble d'appartenance est égale 1 pour tout y. Alors, la similarité de ces qualificateurs, $sim\ (R_1,\ R_2)$, est définie comme le minimum des deux éléments cités ci-dessous:

1. la Similarité $\frac{\int(\mu_{R_1} \cap \mu_{R_2})}{\int(\mu_{R_1} \cup \mu_{R_2})}$ est calculée en utilisant la mesure de Jaccard des ensembles R_1 et R_2 , où μ_{R_1} et μ_{R_2} sont les fonctions d'appartenance des ensembles flous, elle est utilisés pour modéliser les qualificatifs R_1 et R_2 .
2. La similarité entre les valeurs de degré de focalisation pour deux résumés linguistique est calculée comme :
   $1 - \left| d_{foc}(Q1R1\ y's \text{ are } P1) - d_{foc}(Q2R2\ y's \text{ are } P2) \right|$ (3.34)

Par conséquent ,

$$sim\ (R_1,\ R_2) = \min\left(\frac{\int(\mu_{R_1} \cap \mu_{R_2})}{\int(\mu_{R_1} \cup \mu_{R_2})}, 1 - \left| d_{foc}(Q1R1\ y's \text{ are } P1) - d_{foc}(Q2R2\ y's \text{ are } P2) \right|\right) \quad (3.35)$$

Donc, si les deux protoformes sont simples, nous avons $sim\ (R_1,\ R_2) = 1$, puisque les degrés de focalisation sont nuls et les ensembles flous sont l'ensemble d'univers.

Alors, la similarité totale entre deux résumés linguistiques d'une protoforme de type2 est définie comme :

$$sim(Q1R1\ y's \text{ are } P1, Q2R2\ y's \text{ are } P2) = \min\left(sim\ (P_1,\ \mathrm{P}_2), sim\ (Q_1,\ \mathrm{Q}_2), sim\ (T_1,\ \mathrm{T}_2), sim\ (R_1,\ R_2)\right) \quad (3.36)$$

Alternativement, Auteurs Wilbek a proposé la notion de dis-similarité, définie comme suit :

$$d(Q1R1\ y's \text{ are } P1, Q2R2\ y's \text{ are } P2) = 1 - sim(Q1R1\ y's \text{ are } P1, Q2R2\ y's \text{ are } P2) \quad (3.37)$$

NB : la mesure de dis-similarité produit des valeurs dans la plage [0,1].

Il faut prouver que la dis-similarité est considéré comme une distance métrique.

Nous savons que la distance de Jaccard est une métrique.

En plus, dans[119][120], ont proposé cette égalité : $d\ (A,\ B)\ +\ d\ (B,\ C)\ =\ d\ (A,\ C)$ si et seulement si le $max\ (sim\ (A,\ B)+sim\ (B,\ C)\ -\ 1,\ 0)\ \leq sim\ (A\ ,\ C)$.

Alors, cette distance est une métrique.

Supposons que, $d_1\ , d_2\ :\ M\ \times\ M\ \rightarrow\ [0,\ 1]$ sont deux métriques dans la même espace.
Alors $d(x,\ y)=max(d_1(x,\ y),d_2(x,\ y))$ est aussi une métrique.

Donc, nous sommes maintenant prêts à prouver que le résultat principal de dis-similarité de deux résumés linguistiques est une métrique dans l'espace de protoforme des résumés.

Le théorème de dés-similarité de deux résumés linguistiques est**:**

$$d(Q1R1\ y's\ \text{are}\ P1,Q2R2\ y's\ \text{are}\ P2)=1-\min(sim\ (P_1,\ \mathrm{P}_2),sim\ (Q_1,\ \mathrm{Q}_2),sim\ (T_1,\ \mathrm{T}_2),sim\ (R_1,\ R_2))$$
$$=\max(1-sim\ (P_1,\ \mathrm{P}_2),1-sim\ (Q_1,\ \mathrm{Q}_2),1-sim\ (T_1,\ \mathrm{T}_2),1-sim\ (R_1,\ R_2)) \quad (3.38)$$

Pour:

$$1-sim(P_1,P_2)=1-\min\left(\frac{a}{b},\frac{\int(\mu_{P_1}\cap\mu_{P_2})}{\int(\mu_{P_1}\cup\mu_{P_2})}\right)=\max\left(1-\frac{a}{b},1-\frac{\int(\mu_{P_1}\cap\mu_{P_2})}{\int(\mu_{P_1}\cup\mu_{P_2})}\right) \quad (3.39)$$

Donc, $1-\frac{a}{b}\ et\ 1-\frac{\int(\mu_{P_1}\cap\mu_{P_2})}{\int(\mu_{P_1}\cup\mu_{P_2})}$ sont métriques de Jaccard car $(1-\frac{a}{b}$) défini sur l'ensemble des attributs et ($1-\frac{\int(\mu_{P_1}\cap\mu_{P_2})}{\int(\mu_{P_1}\cup\mu_{P_2})}$) défini sur les ensembles flous qui définissent les résumés.

Par conséquent, $1-sim(P_1,P_2)$ est une métrique , selon théorème précédent.

De plus, $1-sim(\mathrm{Q}_1,\mathrm{Q}_2)$ est aussi une métrique.

Et pour $1-sim\ (T_1,\ T_2)=1-1+|\ T_1-T_2|=|T_1-T_2|$, est aussi une métrique.

Il reste à vérifier si le quatrième facteur est aussi une métrique.

Nous avons :

$$1-sim\ (R_1,\ R_2)=1-\min\left(\frac{\int(\mu_{R_1}\cap\mu_{R_2})}{\int(\mu_{R_1}\cup\mu_{R_2})},1-\ \left|\ d_{foc}(Q1R1\ y's\ \text{are}\ P1)-\ d_{foc}(Q2R2\ y's\ \text{are}\ P2)\ \right|\right)$$

$$= \max\left(1-\frac{\int\left(\mu_{R_1}\cap\mu_{R_2}\right)}{\int\left(\mu_{R_1}\cup\mu_{R_2}\right)}, \left| d_{foc}\left(Q1R1\ y's \text{ are } P1\right) - d_{foc}\left(Q2R2\ y's \text{ are } P2\right)\right|\right)$$

(3.40)

Nous savons que $\left(1-\frac{\int\left(\mu_{R_1}\cap\mu_{R_2}\right)}{\int\left(\mu_{R_1}\cup\mu_{R_2}\right)}\right)$ est une distance de Jaccard donc elle est métrique.

Aussi, $\left| d_{foc}\left(Q1R1\ y's \text{ are } P1\right) - d_{foc}\left(Q2R2\ y's \text{ are } P2\right)\right|$ est de la forme 1-norme donc elle est métrique.

Par conséquent, les quatre éléments $1-sim\left(P_1, P_2\right)$, $1-sim\left(Q_1, Q_2\right)$, $1-sim\left(T_1, T_2\right)$, $1-sim\left(R_1, R_2\right)$ sont des mesures et leur maximum est une métrique.

[114]

3.5.3 Similarité entre les ensembles des résumes linguistiques

Il existe plusieurs méthodes pour mesurer la similarité des séries de données, en utilisant les méthodes d'agrégation, les plus connues sont : OWA et Intégral Sugeno.

Dans[121], il a été prouvé qu'on peut utilisé les méthode d'agrégation afin de mesurer la similarité des ensembles de résumés linguistiques.

Tous d'abord, Nous considérons deux ensembles de résumés linguistiques: S1 et S2.

Supposons que l'ensemble S1 contient n résumés: $\{s_{11}, s_{12}, \ldots, s_{1n}\}$ alors que l'ensemble S2 contient m résumé s: $\{s_{21}, s_{22}, \ldots, s_{2m}\}$.

Nous pouvons créer la matrice de similarité de deux résumés s_{1i} et s_{2j} où i = 1, ..., n et j = 1, ..., m.

La forme générale de la matrice de similarité entre ces deux ensembles de résumés est:

$$\begin{array}{c|ccc} & s_{11} & \cdots & s_{1n} \\ \hline s_{21} & sim(s_{11}, s_{21}) & \cdots & sim(s_{1n}, s_{21}) \\ \vdots & \vdots & \vdots & \vdots \\ s_{2m} & sim(s_{11}, s_{2m}) & \cdots & sim(s_{1n}, s_{2m}) \end{array} \qquad (3.41)$$

Alors, il faut trouver le degré de similarité pour chaque résumé . c.à.d. trouver l'ensemble de résumé de S1 le plus similaire par rapport à S2.

Il existe plusieurs travaux qui ont essayé de traiter ce problème :

Au début, la moyenne des valeurs de similarité est calculé par :

$$sim_M(\mathrm{S1},\mathrm{S2}) = \sum_{i=1}^{n} \frac{\max_{j=1\ldots m} sim(s_{1i}, s_{2j})}{n+m} + \sum_{j=1}^{m} \frac{\max_{i=1\ldots n} sim(s_{1i}, s_{2j})}{n+m} \qquad (3.42)$$

Le faible degré de similarité entre deux ensembles de résumés , proposé par [122]ou [123],a un quantificateur des résumés . il a au moins un résumé comparable dans l'autre ensemble. Alors, le faible degré de similarité est :

$$sim_Q(\mathrm{S1},\mathrm{S2}) = \mu_Q\left(\sum_{i=1}^{n} \frac{\max_{j=1\ldots m} sim(s_{1i}, s_{2j})}{n+m} + \sum_{j=1}^{m} \frac{\max_{i=1\ldots n} sim(s_{1i}, s_{2j})}{n+m} \right) \qquad (3.43)$$

Où Q désigne le quantificateur (la plupart ou presque tous) qui est représenté par un ensemble flou approprié sur les entiers non négatifs.

Et **max** correspond au quantificateur (au moins un). [114]

3.5.3.1 Agrégation en utilisant opérateurs OWA :

Un opérateur OWA [124] de dimension n est un mappage.

$F_w : [0,1]^n \to [0,1]$ tel que $W = [w_1, w_2, \ldots w_n]^T$ est un vecteur de pondération avec

$w_i \in [0,1]$ pour tous $i = 1,\ldots,n$, $\sum_{i=1}^{n} w_i = 1$

Et $$F(\mathrm{a}_1, a_2, \ldots, a_n) = W^T B = \sum_{j=1}^{n} w_j b_j \qquad (3.44)$$

Où b_j est le j élément le plus grand de l'ensemble $\{a_1, a_2, \ldots a_n\}$ et $B = [b_1, b_2, \ldots b_n]$.

Pour les quantificateurs normaux non décroissants Q, Yager [27], génère le vecteur de pondération :

$$W = [w_1, w_2, \ldots w_n]^T \qquad (3.35)$$

Où $w_i = \mu_Q\left(\frac{i}{n}\right) - \mu_Q\left(\frac{i-1}{n}\right)$, $i = 1,\ldots,n$.

Par définition, si on a : $\mu_Q(0) = 0$ et $\mu Q(1) = 1$ alors, on aura $w_1 + w_2 + \ldots + w_n = 1$.

Cette procédure de détermination du vecteur de pondération est simple et intuitive. Ainsi, la similarité de deux séries de résumés linguistiques est :

$$sim_{OWA}(S_1, S_2) = \sum_{l=1}^{m+n} w_l b_l \qquad (3.46)$$

Où $w_l = \mu_Q\left(\frac{l}{m+n}\right) - \mu_Q\left(\frac{l-1}{m+n}\right)$ et b_j est le **l** ème élément le plus grand de l'ensemble.

Dans l'ensemble $\{a_1, a_2, \ldots, a_{m+n}\}$:

$$\begin{cases} a_k = \max_{j=1\ldots m} sim(s_{1k}, s_{2j}) & k = 1, \ldots, n \\ a_k = \max_{i=1\ldots n} sim(s_{1i}, s_{2(k-n)}) & k = n+1, \ldots, n+m \end{cases} \quad (3.47)$$

Pour généraliser d'avantage, il faut utiliser les intégrales floues Sugeno ou Choquet afin effectuer la fusion [125].

3.5.3.2 Agrégation de Intégral Sugeno

Soit $X = \{x_1, x_2, \ldots, x_n\}$ un ensemble fini. Alors, la mesure floue [126] sur X est une fonction $g : P(X) \rightarrow [0, 1]$ telle que :

- $g(\varnothing) = 0$
- $g(X) = 1$
- Si $A \subseteq B$ alors $g(A) \leq g(B)$, $\forall A, B \in P(X)$ où P(X) désigne l'ensemble de tous les sous-ensembles de X.

Soit, g est une fonction de mesure floue et h est une fonction $h : X \rightarrow [0, 1]$.

Supposons que les éléments $\{x_i\}$ sont ordonnées de sorte que $h(x_1) \geq h(x_2) \geq \ldots. \geq h(x_n)$.

Alors, l'intégrale discrète Sugeno d'une fonction h par rapport à g est une fonction : $S_g : [0, 1]n \rightarrow [0, 1]$ telle que :

$$S_g = \max_{i=1,\ldots,n} \left[\min \left(h(x_i), g(A_i) \right) \right] \quad (3.48)$$

Où $A_i = \{x_1, x_2, \ldots x_i\}$.

De même, l'intégrale discrète Choquet [127] est une fonction h par rapport à g et elle est désignée par $C_g : [0, 1]^n \rightarrow [0, 1]$ telle que :

$$C_g(h) = \sum_{i=1}^{n} \left[h(x_i) - \mathrm{h}(\mathrm{x}_{i+1}) \right] . g(A_i) \quad (3.49)$$

Où $h(x_{n+1}) = 0$.

Dans notre contexte, $X = S_1 \cup S_2$, c'est-à-dire l'union de l'ensemble des résumés S_1 et S_2.

On a une mesure floue $g : P(X) \rightarrow [0, 1]$ est définie comme :

$g(A) = \mu_Q\left(\frac{|A|}{|X|}\right)$ où valeur absolu|-| dénote la cardinalité de l'ensemble A et X, et μ_Q est la fonction d'appartenance du quantificateur.

Dans notre cas, on a utilisé le quantificateur "le plus".

La fonction de support partiel $h : X \rightarrow [0, 1]$, est définie comme suit :

$$h(s) = \begin{cases} \max_{j=1,\ldots,m} sim(s, s_{2j}) \; si \in S_1 \\ \max_{i=1,\ldots,n} sim(s_{1i}, s) \quad si \in S_2 \end{cases} \qquad (3.50)$$

C'est-à-dire le degré maximal de similarité d'un résumé S est un résumé de l'autre ensemble.

Alors, la valeur de similarité trouvée, en utilisant l'intégrale Sugeno , est exprimée comme suit :

$$S_g(h) = \max_{i=1,\ldots,n}\left[\min(\alpha, \mu_Q\left(\frac{|A_\alpha|}{n+m}\right)\right] \qquad (3.51)$$

Où $|A_\alpha|$ est le nombre de résumés de S_1 et S_2 . La valeur de similarité est de trouvé la plus grande valeur de similarité d'un résumé par rapport au degré de α.

Donc, l'intégrale de Sugeno offre une riche famille de techniques de fusion lorsque les mesures sont modifiées. [121]

3.5.4 Résumés linguistiques différentiels

Les résumés linguistiques enrichis proposés [117]sont composés de deux parties: une partie mettant en évidence les différences entre les sous-ensembles de données avec des étiquettes de catégories différentes et une partie qui fait référence à toutes les étiquettes de catégorie combinées. La protoforme de résumé linguistique différentiel est :

Les différences :

$$Q\ y's\ with\ c_1\ are\ P$$
$$while\ y's\ with\ c_2\ do\ not(d, T) \qquad (3.51)$$

Global :

$$Q\ y's(c_1 \cup c_2)\ are\ P(T) \qquad (3.52)$$

Les résumés linguistiques enrichis ,proposés [117] , visent à rendre les résumés linguistiques plus complets, ceci en combinant les résumés linguistiques globaux classiques (3.52), qui s'appliquent à l'ensemble des données, avec des résumés différentiels (3.51).

Dans l'expression (3.51) , on met en évidence la différence entre les résumés et les différentes classes. La négation " $c_2 do\ not$ " dans le résumé différentiel " $Qy's\ with\ c_1\ areP\ while\ y's\ with\ c_2\ do\ not$ " se réfère à tout le résumé " $Q\ y'\ s$ are P " et pas seulement au quantificateur de résumé.

Exemple : «**la plupart des patients masculins ont une fréquence cardiaque élevée alors que les patients féminins ne l'ont pas**».

La partie différentielle est associée à deux critères d'évaluation:

- **d** indique dans quelle mesure le résumé différencie les deux étiquettes de catégories.
- **T** est le degré de vérité du résumé linguistique " $Q\ y's\ with\ c_1\ are\ P$ ".

Cette double évaluation implique une double condition qui est imposée lors de la sélection des résumés différentiels pertinents .Pour faire partie de la description finale des données, on a besoin deux paramètres de seuil qui sont définis par l'utilisateur, α1 et α2. où $d \geq \alpha_1$ et $T \geq \alpha_2$.

L'expression (3.52) est composée des résumés linguistiques généraux en sous forme de protoforme classique type1, avec une valeur de vérité élevée, et supérieure à un seuil α_3 .

Pour ce travail prenons comme valeur $\alpha_1 = \alpha_2 = \alpha_3 = 0.5$.

Il convient de souligner que les expressions (3.51) et (3.52) sont illustrés par la protoforme le plus simple (3.24). Les résumés différentiels peuvent être basés sur la forme la plus complexe (3.25).

Évaluation proposée du degré de vérité :

L'évaluation des résumés différentiels est basée sur les degrés de vérité et le critère différentiel.

Ces critère comparent $LS_1 = Qy\ 's\ with\ c_1\ are\ P$ et $LS_2 = Qy'\ s with\ c_2$ are P , où ces deux résumés LS1 et LS2 ont le même quantificateur, le même quantificateur temporel, le même résumer et le même qualificateur, mais les étiquettes de catégories sont différentes. Après cela on déduit qu'un des deux résumés linguistique doit avoir un degré de vérité élevé et l'autre avec un degré faible. Donc, le critère différentiel est :

$$d = | T_1 - T_2 | \qquad (3.53)$$

Si deux résumé sont différent légèrement par leur quantificateur .Alors leur résumer sont opposés l'un à l'autre. Les différentiels $LS_1 = Qy\ 's\ with\ c_1\ are\ P_1$ et $LS_2 = Qy'\ s with\ c_2$ are P_2 sont similaire et Q_1 et Q_2 le sont aussi.

Dans le cas général, nous avons :

$$d = d(LS_1, LS_2)cmp(c_1, c_2) \qquad (3.54)$$

Où $d\left(LS_1,\ LS_1\right)$ est la mesure de dis-similarité, proposée par Wilbik, qui est appliquée aux résumés linguistiques et $cmp(c_1, c_2)$ est une mesure de comparaison pour les étiquettes de catégorie définies par : $cmp(c_1, c_2) = \begin{cases} 1 & si\ c_1 \neq c_2 \\ 0 & \text{sinon} \end{cases}$

Donc, si les résumés considérés LS1 et LS2 sont applicables à la même étiquette de catégorie, Alors, ils sont associés à $cmp(c_1, c_2) = 0$, $d = 0$ et ils ne satisfont pas la condition sur le critère différentiel minimal.

Dans ce cas, les critères de résumé sont identiques, à l'exception de leurs étiquettes catégorielles, Donc, on a : $sim\left(P1,\ P2\right) = 1, sim\left(Q1,\ Q2\right) = 1$

$sim(R_1, R_2) = 1$, $sim(\mathrm{Q}_{\tau 1}, \mathrm{Q}_{\tau 2}) = 1$ et $cmp(c_1, c_2) = 1$

Par la réduction : $d = (1 - sim(\mathrm{LS}_1, LS_2))cmp(c_1, c_2)$

$$= (1 - \min(1,1,1, sim(T_1, T_2), 1))1$$

$$= \left|T_1 - T_2\right| \text{ [117]}$$

3.5.5 Travaux de littératures

Nous citons par la suite quelques travaux qui ont utilisé le calcul de similarité entre les résumés linguistiques.

- **Dans le travail**[121] :

Ils ont développé un système de calcul métrique entre les résumés, basés sur les protoformes de Yager et les méthodes d'agrégation.

Ils ont fournis une étude de cas de soins des personnes âgées à TigerPlace (établissement de «vieillissement en place» à Columbia, au Missouri).[128].Dans ce travail, ils ont créé les résumés linguistiques pour un patient de sexe masculin, âgé d'environ 80 ans. Ces données

proviennent de deux capteurs: l'agitation du lit qui illustre le mouvement en position allongée et le mouvement de la chambre pendant une nuit (de 21h à 7h).

Au début, ils ont crée un ensemble des résumés linguistique à partir d'une base donnée, ces résumé linguistique sont présentés par la protoforme :

$Qy's\ are\ P$.

Alors, l'objet de ce travail est de développer un système qui compare les différents mouvements nocturnes d'un patient afin de détecter les changements pendant une nuit.

Ils ont représenté les ensembles flous par la fonction d'appartenance trapézoïdale, par trois valeurs linguistiques: « low , medium ,high)) où ils ont élaboré des ensembles de résumés de données sur deux ou plusieurs fenêtres de temps (deux nuits différentes).

Ensuite, ils ont utilisé le calcul de similarité / -dis similarité [114]sur des paires de résumés examinés .

A partir de ces calculs, ils ont obtenu une matrice de similarité des résumés linguistiques pour les deux nuits énumérées. Pour générer cette matrice, ils ont utilisé les méthodes d'agrégation, pour faire la distinction entre les «bonnes nuits» et les «mauvaises nuits».

Dans l' étude de [117] **:**
Il a été développer un système de calcul capable de générer automatiquement des descriptions linguistiques d'une série chronologique de patient qui sont atteints de choc septique. Ils ont utilisé les information de la base de données MEDAN[129]. Cette base est composée des patients atteints de choc septique abdominal dans unité de soins intensifs (USI) de 70 hôpitaux différents en Allemagne, entre 1998 et 2002.

Dans cette étude, les expériences sont réalisées dans un sous-groupe de 383patients qui répondent aux critères du choc septique abdominal. Le principe de cette étude est de détecté l'état du patient (vivant ou décédé). Cette étude a était évaluée pendant 24 heures, du moment donné jusqu'à ce que le patient soit sorti de l'USI ou décédé.

Pour régler ce problème, ils ont proposons un nouveau type de résumé différentiel, basé sur un critère numérique qui est sous forme d'une contextualisation temporelle et catégorielle.

Alors, ces résumés linguistiques contiennent des observations de la fréquence cardiaque (FC) et les valeurs du test sanguin partiel de la thromboplastine (PTT). Les résultats fournissent un aperçu général des différences entre toutes les observations mesurées de la fréquence dans une période de temps.

Dans l' étude de [130]**:**
Le but du travail proposé par [130] était de produire un système de classification supervisé , en utilisant les résumés linguistiques c.à.d. de classer un vecteur de résumé linguistique ,au lieu de classer une base de données complet. les auteurs ont utilisé les méthodes de

classification classique (SVM,KP ….).Ils ont représenté les séries temporelles par des résumés linguistiques sous forme d'une protoforme proposée par Yager. Le vecteur d'entré contient les variable linguistique qui sont associé par leur degré de vérité.

Parmi les nombreuses approches du problème général de la classification des données, il a été utilisé la méthode Machine à vecteurs de support (SVM) et la méthode le voisin le plus proche de la distance euclidienne.

Les résultats expérimentaux sont présentés pour les séries de données temporelles artificielles et réelles. Ils illustrent les performances de classification proposée avec les résumés linguistiques (CLS).La précision de la méthode proposée est comparée avec des méthodes (SVM, kNN) Mais dans le jeu de données réel, la méthode proposée dans ce travail considéré est encore assez moyenne parmi les autres méthodes.

3.6 Conclusion

Dans ce chapitre, nous avons présenté les principaux concepts de fouille de données (data mining).Nous détaillons l'une de méthode de data mining qui est la classification. Suite à état de l'art présenté de ce chapitre, Le problème major rencontré dans les systèmes de classification supervisé c'est la quantité de données entrées et la vitesse d'exécution de ces données. Donc, il faut un modèle qui peut combiner entre traitement de volume de données et la performance et la rapidité de la classification supervisé.

Aussi dans ce chapitre, Nous avons citées des travaux qu'ils ont réalisé des systèmes prédictifs à partir aux résumés linguistiques de données entrées et des travaux qui ont utilisé la notion de calcul de similarité entre les résumés linguistiques. En effet, la notion de la similarité entre les résumés linguistiques, dans le résumé quantifié (résumé de Yager), est calculée juste pour supprimer les redondances (plusieurs résumés ayant même poids) et les informations sans poids .

Dans le chapitre suivant, nous proposons un système de diagnostic médical qui traite le problème de volume de donnée médical. Notre modèle médical utilise les résumés linguistiques à base de calcul de la cardinalité floue afin de construire une base de connaissances interprétables en langage naturel. Nous abordons sur les performances de notre modèle médical qui est relie les résumés linguistiques avec le système d'interrogation et le classifieur supervisé médical.

chapitre 4. Conception d'un modèle médical à base de résumés linguistiques

chapitre 4. Conception d'un modèle médical à base de résumé linguistique 75

4.1 Introduction 76

4.2 Matériels et méthodes 76

4.2.1 Langages et outils utilisés 76
4.2.2 Les bases de données 77
4.2.3 Critères d'évaluations 78

4.3 Architecture d'un nouveau modèle de diagnostic médical 79

- Aperçue sur l'application 80

4.4 Approche proposée de résumé linguistique médical RLR-CardF 81

4.4.1 Implémentation de RLR-CardF sur une base de données médicale 84
4.4.2 Résultats et discussions 86
4.4.3 Aperçue sur l'application 88

4.5 Approche proposée d'un système d'interrogation médical flexible 90

4.5.1 Calcul de quantificateur flou par le produit scalaire 90
4.5.2 Calcul le quantificateur flou par le degré de vérité 92
4.5.3 L'approche de recherche proposé IFlex- RLR-CardF ,basé sur le calcul le degré de validité 93
4.5.4 Aperçue sur l'application 97

4.6 Approche proposé d'un classifieur médical « Classifieur- RLR-CardF»: 98

4.6.1 Schéma général de approche proposé « Classifieur- RLR-CardF » 98
4.6.2 Processus de développement, Classifieur- RLR-CardF 99
4.6.3 Aperçue sur l'application 104

4.7 Résultat et discussion 105

4.8 Comparaison de l'approche proposée avec les travaux de la littérature 111

4.9 Conclusion 112

4.1 Introduction

Les systèmes d'aide au diagnostic médial classique traitent des bases de données numériques .Mais, la capacité de ces bases de données s'augmentent de plus en plus ce qui cause un problème au niveau de performance de ces systèmes. Dans le domaine médical, il est important de crée un système de décision robuste, rapide et fiable ; afin de détecter une pathologie de données.

Dans les chapitres précédents, nous avons dressé une étude bibliographique concernant les problèmes de volume important des données numériques médicaux, l'intreprétabilité de ces données et les performances des classifieurs . Nous avons citées quelque méthodes et approches de résolution qui leur ont été dédiées.

Dans cette partie de notre thèse, nous intéressons à la conception d'un modèle de diagnostic médical complet qui combine entre trois approches qui sont proposées dans notre travail de recherche :

RLR-CardF(**R**ésumé **L**inguistique **R**éduit à base de **c**alcul du **Card**inalité **F**loue).

IFlex- RL (**I**nterrogation **Flex**ible des **R**ésumés **L**inguistiques).

Classifieur- RLR-CardF(**Classifieur** à base **RLR-CardF**).

Ainsi, nous présentons, ci-dessous, les descriptions de ces trois approches :

RLR-CardF : Représente notre méthode de résumé linguistique, à base de calcul de cardinalité floue, afin de faire une réduction sémantique de la base de données médicale.

IFlex- RL : Représente notre algorithme d'interrogation flexible afin d'interroger nos résumés linguistiques des données médicales et chercher des réponses exactes des requêtes, selon les préférences de l'expert du domaine.

Classifieur- RLR-CardF: A l'aide de calculs de la similarité, entre les résumés linguistiques, nous proposons d'adapter ces calculs pour créer un modèle de diagnostic médical, c. à. d, nous proposons une nouvelle méthode de la classification supervisée.

Afin de valider nos approches et évaluer les performances de nos modèles proposées, nous allons utiliser trois bases de données médicales (PIMA,WBCD, Mamography).

4.2 Matériels et méthodes

4.2.1 Langages et outils utilisés

Pour l'implémentation de notre architecture nous avons utilisé les outils et les langages suivants :

- Nous avons réalisé l'étude expérimentale sur un ordinateur avec processeur Dual-Core 2,20GHz.

- *Le langage java*, pour le développement le résumé à base de calcul de la cardinalité floue, la construction de système d'interrogation flexible et le système de classification . On a travaillé sous l'environnement du développement *NetBeans (NetBeans IDE 8.0 for Java EE Developers)*

4.2.2 Les bases de données

Dans cette thèse de doctorat, nous utilisons trois bases de données médicales extraites du dépôt d'UCI (A. Frank and A. Asuncion) [131].

Les bases de données sont citées dans le tableau ci-dessous : Pima Indian Diabetes (PIMA), Wisconsin Breast Cancer (WBC), et mammographie (Mamography).

Tableau 4-1 Caractéristiques des jeux de données utilisés

Jeux de données	Attribut	Classe	Instances
PIMA	8	2	768
WBCD	9	2	683
Mamography	5	2	961

Ces jeux de données ont été largement testés dans les travaux des systèmes de classification, nous présentons ci-dessous les descriptions de ces ensembles :

4.2.2.1 PIMA

L'ensemble de données médical ''Indiana Pima du Diabètes '' a été choisi à partir de dépôt d'UCI(A. Frank and A. Asuncion) où on traite une étude sur 768 femmes Indiennes. Le diagnostic est une valeur binaire variable «classe» qui permet de savoir si le patient montre des signes de diabète selon les critères de l'O.M.S. Les exemples de données sont répartie en deux classe ,500 exemples appartiennent à la classe 0 (Classe non diabétique) et 268 exemples appartiennent à la classe 1 (Classe diabétique)). Les huit descripteurs cliniques sont :

1. Npreg : nombre de grossesses (Ngross).
2. Glu : concentration du glucose plasmatique (mg/dl).
3. BP : tension artérielle diastolique (mm Hg) (PAD).
4. SKIN : épaisseur de pli de peau du triceps (mm).
5. Insuline : dose d'insuline (mu U/ml).
6. BMI : index de masse corporelle (poids en kg/ m^2).
7. PED : fonction de pedigree de diabète (l'hérédité).

4.2.2.2 Wisconsin Breast Cancer (WBCD)

L'ensemble de données du cancer du sein dénommé « Wisconsin Breast Cancer Dataset » a été collecté à l'Université du Wisconsin. Les mesures sont assignées à une valeur entière comprise entre 1 et 10 (1 étant la plus proche de bénigne et 10 la plus proche de maligne).

Après élimination des données manquantes dans cette base de données .On obtient un ensemble de 683 patientes. Ces exemples sont répartie en deux classe (546 exemples

appartiennent à la classe 2 (Classe bénigne), 137 exemples appartiennent à la classe 4 (Classe maligne)). Il contient 699 échantillons avec neuf descripteurs qui sont :

1. Clump Thickness
2. Uniformity of Cell Size
3. Uniformity of Cell Shape
4. Marginal Adhesion
5. Single Epithelial Cell Size
6. Bare Nuclei
7. Bland Chromatin
8. Normal Nucleoli
9. Mitoses

4.2.2.3 Mammographie

Les données de cette base sont des données de mammographie Bénin ou Maligne. L'ensemble de ces données sera utilisé pour prédire le cancer du sein et la gravité d'une lésion de la masse mammographie. Cette base est évaluée sur plusieurs caractéristiques, parmi ces caractéristiques nous comptons l'âge du patient. Parmi les données recueillis de l'Institut de Radiologie de l'Université d'Erlangen-Nuremberg entre 2003 et 2006, 516 sont bénin et 445 masses sont malignes. Les huit descripteurs cliniques sont :

1. BI-RADS assessment: (ordinal, non-predictive)
2. Age: patient' age
3. Shape: mass shape: (round, oval, lobular, irregular)
4. Margin: mass margin
5. Density: mass density
6. Severity: Benign, malignant

4.2.3 Critères d'évaluation

Afin de tester les performances de notre classifieur, nous avons divisé la base de données en deux sous-ensembles : 4/5 pour l'apprentissage et 1/5 pour le test. Par la suite, l'évaluation est donnée selon le taux de la classification (**TC**), la sensibilité (**Se**) et la spécificité (**Sp**).

Dans notre thèse, il nous faut utiliser trois critères d'évaluations pour tester la capacité de prédiction des modèles construits, les critères sont définis comme suit :

- ✓ **Le taux de classification correcte (TC%)** (le taux de reconnaissance) , il est calculé par :

$$TC = 100 * \frac{VP + VN}{VN + VP + FN + FP}$$

Où VP, VN, FP et FN désignent respectivement : vrais positifs, vrais négatifs, faux positifs et faux négatifs.

- VP : la classification correcte des échantillons positifs.
- VN : la classification correcte des échantillons négatifs.
- FP : la classification erronée des échantillons négatifs dans les échantillons positifs.
- FN : la classification erronée des échantillons positifs dans les échantillons négatifs.

- ✓ **Sensibilité (SE%)** : c'est le pourcentage d'échantillons positifs qui sont correctement classés, il est défini par :

$$SE = 100 * \frac{VP}{VP + FN} .$$

- ✓ **Spécificité (SP%) :** c'est le pourcentage d'échantillons négatifs qui sont correctement classés, il est défini par :

$$SP = 100 * \frac{VN}{VN + FP} .$$

4.3 Architecture d'un nouveau modèle de diagnostic médical

L'objectif principal de notre thèse est de développer un modèle d'aide au diagnostic médical pour des bases de données volumineuses. Donc, il faut chercher une méthode pour condenser les données sémantiquement, sans perde la sémantique des informations importantes.

Il est important que les informations stockées dans notre base de connaissance soient utilisé avec des termes proches du langage naturel. Aussi, il est possible d'effectuer des recherches avec des requêtes en langage humain. C'est à dire de créer un système d'interrogation médical qui est flexible, rapide et précis.

Dans cette thèse, nous proposons un modèle qui réuni quatre grands champs de recherches en biomédicale. Notre modèle traite :

- Problème du volume des bases des données médicales.
- L'interprétabilté des enregistrements médicaux en langage naturel.
- Recherche des informations pertinentes sémantiquement proche au langage naturel.
- Prédiction à partir des résumés sémantiques des données.

Nous schématisons notre travail par la (figure 4.1) qui explique les différentes étapes proposées :

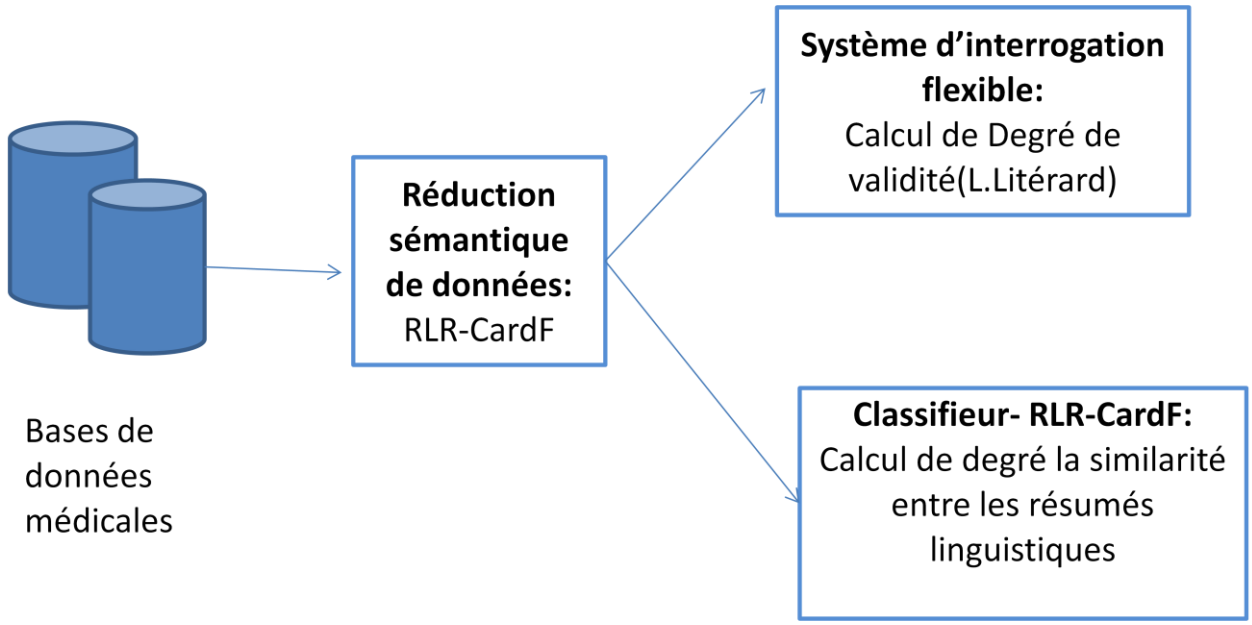


Figure 4-1 schéma général de notre modèle

Dans cette thèse, nous contribuons trois approches qui sont définies ci-dessous :

RLR-CardF(**R**ésumé **L**inguistique **R**éduit à base de **c**alcul de la **Card**inalité **F**loue). : Nous proposons à utiliser un nouvel algorithme de la méthode de résumé linguistique, à base de calcul de cardinalité floue, afin de faire une réduction sémantique de la base de données médicale.

IFlex- RLR-CardF(**I**nterrogation **Flex**ible des **R**ésumés **L**inguistiques à base de calcul de cardinalité floue) : nous proposons un algorithme interrogation flexible qui est basé sur le calcul de degré de validité , afin d'interroger nos résumés linguistiques médicaux et chercher les réponses exacte pour des requêtes sémantiques, selon les préférences de l'expert du domaine.

Classifieur- RLR-CardF (**Classifieur** à base de **RLR-CardF**).: A l'aide de calculs de la similarité, entre les résumés linguistiques, nous proposons d'adapter ces calculs pour créer une nouvelle approches de diagnostic médical, c. à. d, proposons une nouvelle méthode de la classification supervisé.

Dans les prochaines sections, nous détaillerons les différentes approches proposées afin de créer un modèle de diagnostic et de recherche fiable et performant.

- **Aperçue sur l'application**

La figure suivante montre un schéma général de l'application de notre modèle de diagnostic médical :

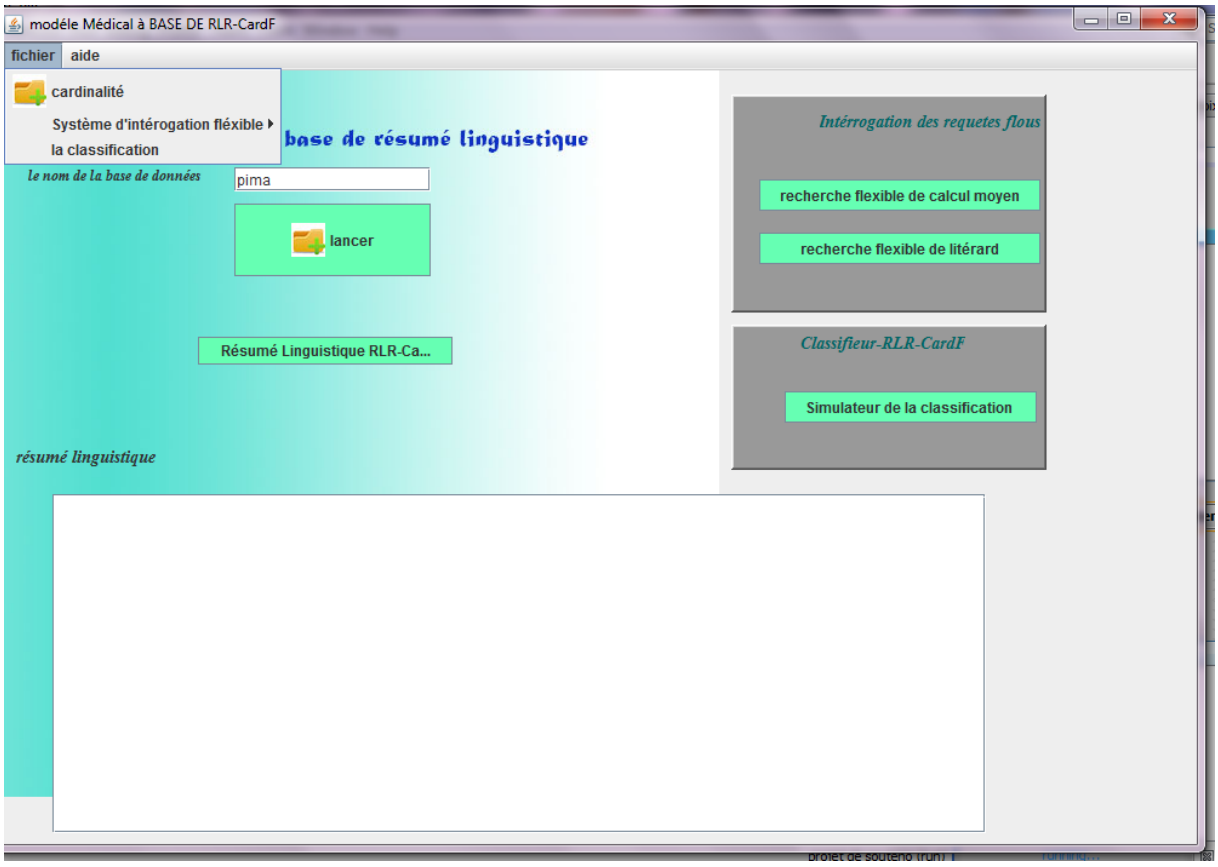


Figure 4-2Fenêtre Principale de notre application médicale

4.4 Approche proposée des résumés linguistiques médicaux RLR-CardF

Nous commençons cette section par la première partie de notre modèle qui est le résumé linguistique à base de calcul de cardinalité floue. L'intérêt de cette méthode est de condenser les informations sémantiquement avec l'enregistrement de tous les degrés et les poids essentiels.

Nous choisissons le résumé linguistique à base de calcul de la cardinalité floue qui contient quatre critères de validités importantes pour la déduction d'une pathologie et la vérification d'une requête.

Ces quatre critères sont :

- Résumé **P** : valeur linguistique d'un attribut (prédicat flou) définie dans le domaine de l'attribut, exemple : jeune pour l'attribut Age.
- Quantité **Q** : quantificateur linguistique, exemple : la plupart.
- Vérité (validité) **T** du résumé : nombre qui évalue le degré vérité (validité) d'un résumé, il appartenant à l'intervalle [0,1], par exemple, 0,7.
- Qualificateur **R** (facultative): deuxième variable linguistique d'un autre attribut, c'est une valeur linguistique (prédicat flou) définie sur le domaine de l'attribut.

En générale, le résumé à base de calcul de cardinalité floue est une approche présentée dans[4] et se développe en deux étapes :

- La première phase : C'est l'étiquetage (fuzzification) où nous remplaçons les enregistrements (n-uplets) par des variables linguistiques.
- La deuxième phase : Nous commençons à calculer la cardinalité flou pour chaque variable linguistique. puis, nous calculons la cardinalité d'intersection entre deux variables linguistiques, jusqu'a l'intersection entre n variable linguistiques.

L'algorithme4.1 résume les différentes étapes de la création de résumé linguistique à base de calcul de cardinalité flou, proposé par Pard Bosc et Dubois :

Algorithme4.1 : la création de résumé linguistique à base de calcul de cardinalité floue (RL-CardF) [4]

Input : n attribut, m $t-uplet_i(a_i,\ldots,a_{in})$.

Out put : linguistics summaries r_{sum}

Begin

Calculated a fuzzification of m $t-uplet_i(a_i,\ldots,a_{in})$

$P=2^n$.

F_{A_r} = 1/0.

where p : number of linguistic variable

A_r : linguistic variable

For j=0 **to** P **then**

Calculate the fuzzy cardinality for each linguistic variable

$$F_{A_j}=1/0+\ldots+1/n-1+1/n+\lambda_1/(n+1)+\ldots+\lambda_k/(n+k)+0/(n+k+1)+\ldots$$

Endfor.

Where: $1>\lambda_1\geq\ldots\geq\lambda_k>\lambda_{k+1}=0$ **and** $n\geq 0$, $k\geq 0$

Calculate intersections between linguistic variables

$$F_{A_{r1}},F_{A_{r2}},\ldots,F_{A_{rp}} \text{ et } (F_{A_{r1}}\wedge F_{A_{r2}}),\ldots,(F_{A_{r1}}\wedge F_{A_{rn}}),\ldots,(F_{A_{r1}}\wedge F_{A_{r2}}\wedge\ldots\wedge F_{A_{rn}})$$

End.

Dans cette algorithme 4.1(**RL-CardF)** , nous remarquons que les calculs de cardinalités floues sont calculés pour toutes les variables linguistiques des attributs(calcul de $F_{A_{r1}},F_{A_{r2}},\ldots,F_{A_{rp}}$).ainsi, les calculs de cardinalité floues pour les m intersections entre les variables linguistiques :calcul de $(F_{A_{r1}}\wedge F_{A_{r2}}),\ldots,(F_{A_{r1}}\wedge F_{A_{rn}}),\ldots,(F_{A_{r1}}\wedge F_{A_{r2}}\wedge\ldots\wedge F_{A_{rn}})$.

Nous voyons que les résumés **RL-CardF** ,proposés par P.Bosc et Dubois[4], contiennent beaucoup d'informations (connaissances). Certaine d'entre elles sont non utile parce que ces résumés linguistiques contiennent des informations avec des poids faibles.

Dans cette phase, notre contribution est de minimiser nombre des résumés linguistiques **RL-CardF** proposé par P.Bosc et Dubois où nous supprimons les informations (les résumés linguistiques) non utile dans le système de déduction. Car, nous remarquons dans la description des résumés linguistiques **RL-CardF**, que l'intersection entre plus de deux variables linguistiques, contiennent des degrés d'appartenances faibles (moins de 0.3).

Dans notre thèse, nous proposons un résumé linguistique réduit, à base de calcul de cardinalité floue, **RLR-CardF** . Il suffit de calculer la cardinalité floue(présenté les résumés linguistiques) pour chaque variable linguistique et les résumé linguistiques de l'intersection entre deux variables linguistiques. Donc, il suffit de calculer $F_{a_1}, F_{a2}, ..., F_{a_n}$ et $(F_{a_1} \wedge F_{a2}), ..., (F_{a_{n-1}} \wedge F_{a_n})$.

Nous remarquons que les résumés linguistiques, calculés par notre algorithme, sont les seules informations importantes pour la prise de décision médicale.

Nous résumons notre résumé linguistique médical **RLR-CardF** par l'algorithme4.2 suivant :

Algorithme4.2 : le résumé linguistique réduit à base de calcul de cardinalité floue **RLR-CardF**

Input : n attribut, m $t-uplet_i(a_i,\ldots,a_{in})$.

Out put : linguistics summaries r_{sum}

Begin

Calculated a fuzzification of m $t-uplet_i(a_i,\ldots,a_{in})$

$P = 2^n$.

F_{A_r} = 1/0.

where p : number of linguistic variable

A_r : linguistic variable

For j=0 **to** P **then**

Calculate the fuzzy cardinality for each linguistic variable

$$F_{A_{rj}} = 1/0 + ... + 1/n-1 + 1/n + \lambda_1/(n+1) + ... + \lambda_k/(n+k) + 0/(n+k+1) + ...$$

Endfor.

Where: $1 > \lambda_1 \geq ... \geq \lambda_k > \lambda_{k+1} = 0$ **and** $n \geq 0$, $k \geq 0$

Calculate intersections between linguistic variables

$F_{A_{r1}}, F_{A_{r2}}, ..., F_{A_{rp}}$ et $(F_{A_1} \wedge F_{A_2}), ..., (F_{A_1} \wedge F_{A_n})$

End.

Afin de faciliter l'interprétation de notre algorithme amélioré **RLR-CardF** (algorithme4.2).Nous l'illustrons notre méthode de résumé linguistique sur une base de données médical PIMA Indiana diabète, afin d'expliquer les étapes de la création d'un résumé linguistique médical amélioré, à base de calcul de la cardinalité floue.

4.4.1 Implémentation de RLR-CardF sur une base de données médicale

Soit r_{su} est un résumé linguistique introduit par le calcul de cardinalité floue RLR-CardF. Il se compose de deux parties (citée précédemment):

La première partie : Nous impliquons le processus de la fuzzification qu'il consiste à transformer les valeurs réelles en variables linguistiques qui sont associé à des dégrées d'appartenances.

Cependant, chaque attribut a été composé de plusieurs sous-ensembles flous, selon les experts dans le domaine.

Dans notre modèle, nous décomposons chaque attribut en deux sous ensembles flous, pour rendre la partie de fuzzification automatique.

Soit, $A_{r_1},\ldots,A_{r_n}$ sont des ensembles flous définis par des fonctions d'appartenances,

Où $\mu_{A_{ri}}(a_i)$ est le degré d'appartenance dans un ensemble flou qui est exprimé par un nombre compris entre 0 et 1.

$$\text{Où } \begin{matrix}\mu_A : \Omega \to [0;1] \\ x \to \mu_A(x)\end{matrix} \qquad (4.1)$$

Nous utilisons la fonction d'appartenance de type trapèze (figure4.2). Cette fonction est définie par : une limite inférieure ***a,*** une limite supérieure ***d***, un support inférieur limite ***b*** et un support supérieur limite ***c***.

Où ***a <b <c <d***:

$$\mu_A(x) = \begin{cases} 0 & si(x < a)\,\text{ou}(\text{x} > \text{d}) \\ \dfrac{x-a}{b-a} & si(a \le x \le b) \\ \dfrac{d-x}{d-c} & si(b \le x \le c) \\ & \text{si}(\text{c} \le \text{x} \le \text{d}) \end{cases} \qquad (4.2)$$

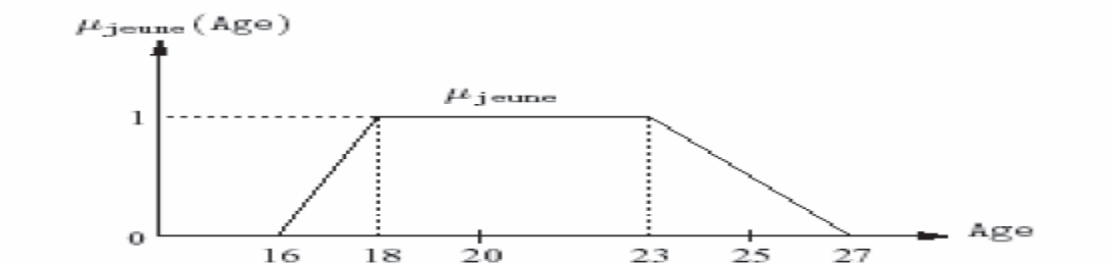


Figure 4-3Représentation d'un sous-ensemble fou Jeune

Pour rendre la phase de fuzzification automatique, nous utilisons deux variables linguistiques par exemple :(normal et élevé) et des fonctions d'appartenances trapézoïdales.

Pour cela nous appliquons la fuzzification dans la Base de données ''Pima Indiana diabète'' où nous composons chaque attribut en deux sous ensembles flous qui sont représentés par les schémas suivants :

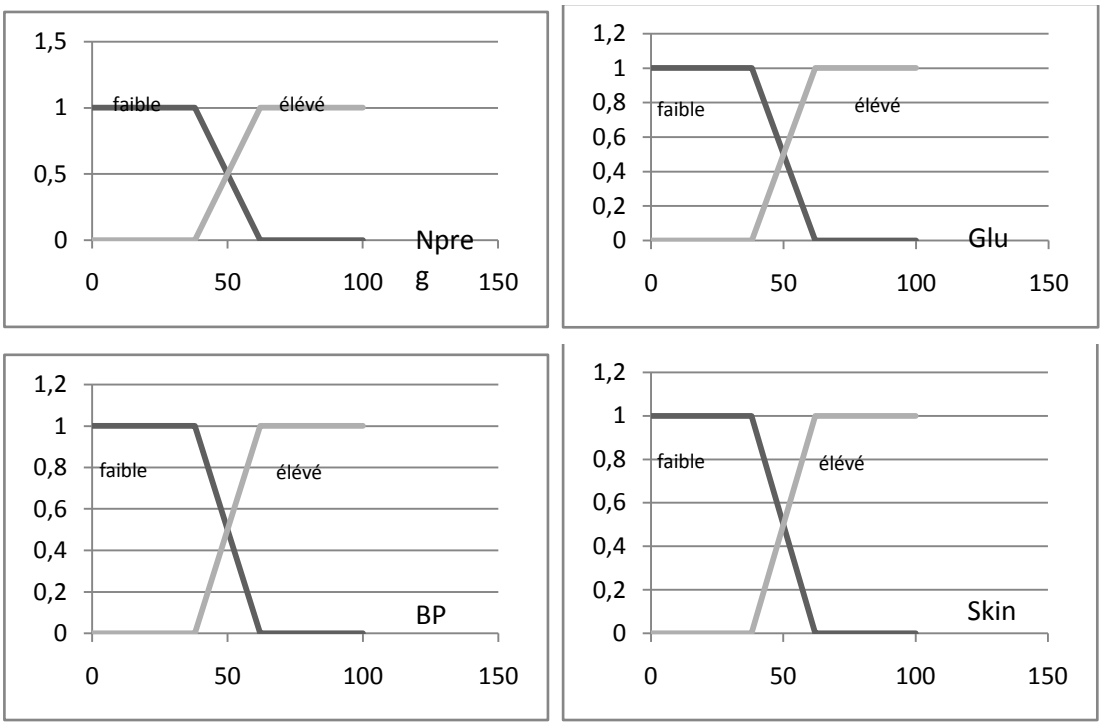


Figure 4-4A partition floue des attributs Npreg,glu,Bp,SKIN

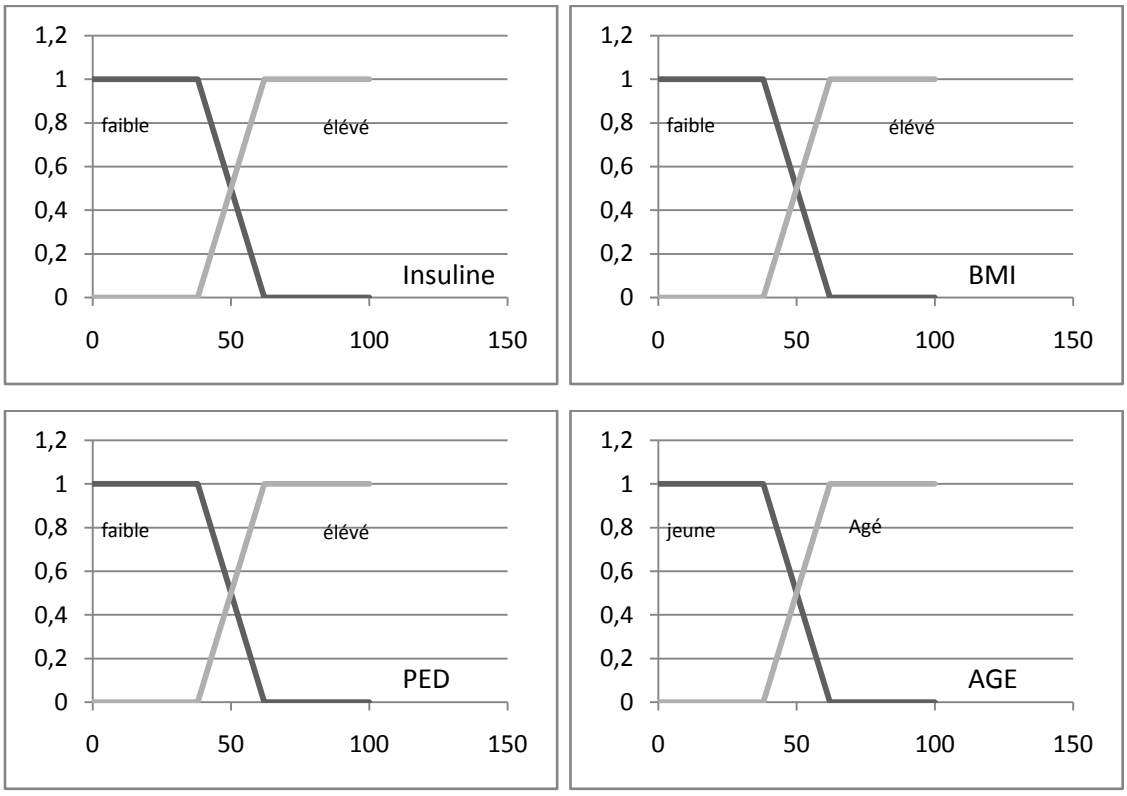


Figure 4-5A partition floue des attributs insulin,BMI,PED,Age

En résumé, la fuzzification est une étape qui consiste à transformer les valeurs numériques en variables linguistiques. Nous décomposons chaque attribut à deux sous-ensembles flous d'une façon automatique. Après la fuzzification et le choix des variables linguistiques, nous calculons les cardinalité floues de $\alpha - cut$ pour chaque variable linguistique et l'intersection entre eux. Nous détaillons nos résultats obtenus dans la sous section suivante.

4.4.2 Résultats et discutions

Selon notre algorithme proposé **RLR-CardF ,** nous élaborons les résultats obtenus de la base de données médicale(PIMA).

Dans l'étape de fuzzification, nous obtenons les résultats des variables linguistiques de la base de données(PIMA) qui sont associées par des degrés d'appartenances (tableau4. 2).

Sachant que μ_A est une fonction d'appartenance de l'ensemble flou A.

$\forall x \in X \ \ \mu_A \in [0,1]$ Ensemble A est défini par:

$$A = \{(x, \mu_A(x)) | x \in X\} \qquad (4.3)$$

Tableau 4-2 Tableau dela fuzzification de la base de données PIMA

Patients	**Nbre gr**		**GLU**		**BMI**		**...**	**PED**		**AGE**		**Classe**	
1	***faible***	***élevé***	***Faible***	***élevé***	***faible***	***Elevé***	***...***	***Faible***	***élevé***	***jeune***	***Agé***	***Non malade***	***Malade***
2	1	0	1	0	1	0	...	1	0	1	0	1	0
3	1	0	0.3	0.7	0.6	0.4	...	0	1	1	0	0	1
4	1	0	1	0	0.9	0.1	...	1	0	1	0	0	1
	⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮	⋮
294	1	0	0.7	0.3	1	0	...	1	0	1	0		1

Dans nos résumés produits, la notion d'utilisation de la logique floue est très intéressante parce qu'elle fournit une formulation linguistique en langage humain. En outre, l'utilisateur n'a pas une idée précise de ce qui peut être obtenu à partir des données stockées, c'est pour cela que la théorie des ensembles flous correspond vraiment à cette exigence.

La deuxième partie : il consiste à créer les résumés linguistiques en utilisant l'algorithme 4.2 de calcul de la cardinalité floue **RLR-CardF**.

Selon l'algorithme4.2, nous illustrons le calcul de cardinalité pour chaque variable linguistique de la base de données(BD) .Nous obtenons les résumés linguistiques (**RLR-CardF** de la **BDD -PIMA**) qui sont enregistrés, dans notre système médical, sous forme d'un

fichier texte nommé : **« resumé-pima.Txt »** . Les résumés linguistiques sont illustrés par (figure 4.6).

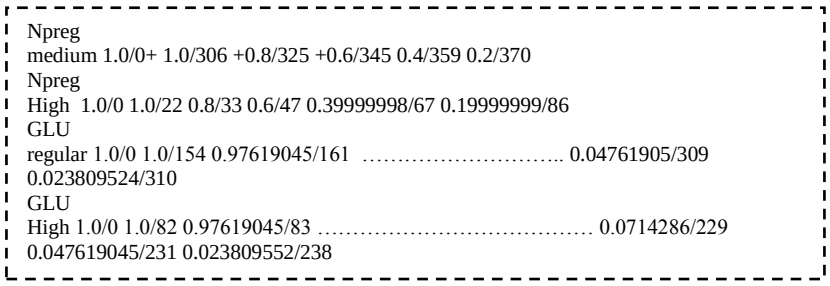


Figure 4-6 Représentation de fichier resumé-pima.Txt (RLR-CardF de la BD -PIMA)

Nos résumés linguistiques contiennent toutes les informations (qualitatives et quantitatives), essentielles qui peuvent être utilisé dans d'autres systèmes d'informatiques.

Alors, cette méthode minimise le taux de perte connu dans des autres méthodes de résumé linguistique. Elle garantit l'information quantitative (cardinalité floue) et l'information qualitative (degré d'appartenance dans le sous-ensemble flou) et elle est interprétable en langage naturel.

Pour facilité la compréhension de nos résumés linguistiques, nous présentons quelque exemples de résumé linguistiques simple (un variable linguistique) et le résumé linguistique complexe (l'intersection entre deux variable linguistique) pour expliquer comment ces résumés linguistiques sont interprétable en langage naturel :

- Selon la figure 4.6, le résumé linguistique de la variable linguistique (faible) de l'attribut (Nombre de grossesse) $F_{Nreg_{low}}$ est:

  $F_{Nreg_{low}}$ = 1,0 / 0 + 1,0 / 306 + 0,8 / 325 + 0,6 / 345 0,4 / 359 0,2 / 370

  Cette écriture est interprétable en langage naturel. Car, le nombre 1.0 (respectivement 0.8, 0.6, 0.4, 0.2) est un degré d'appartenance de sous ensemble flou « faible (low)» de l'attribut « Npreg»(Nombre de grossesse) . Aussi, le nombre 306 (respectivement 325, 345, 359.370 représente le nombre de patients appartenant à la variable linguistique« faible (low)» de l'attribut « Npreg»(Nombre de grossesse) .

  Nous pouvons dire que dans la base de données PIMA

  - ✓ il y a 325 patients qui appartiennent totalement à la variable linguistique« faible(low)» de l'attribut « Npreg»(Nombre de grossesse) .
  - ✓ il y a 19 (325-306 = 19) patients qui appartiennent partiellement à la variable linguistique« faible (low)» de l'attribut « Npreg»(Nombre de grossesse) qui est associé à un degré d'appartenance = 0,8.

- Selon la figure 4.6, le résumé linguistique complexe est :

$$F_{Nreg_{low} \wedge GLU_{High}} = 1.0/0 + 0.8/14 + 0.6/26 + 0.4/37 + 0.2/42$$

Nous pouvons dire qu'il ya :

- ✓ il y a 18 patients qui appartiennent totalement à l'intersection entre la variable linguistique« faible(low)» de l'attribut « Npreg»(Nombre de grossesse) et la variable linguistique« élevé(high)» de l'attribut « GLU»

- Prenons un autre exemple dans les résumés linguistiques **RLR-CardF** de la base de données WBCD. le résumé linguistique de la variable linguistique « Low» (faible) de l'attribut « $UCShape$ »(Uniformity of cell shape) est :

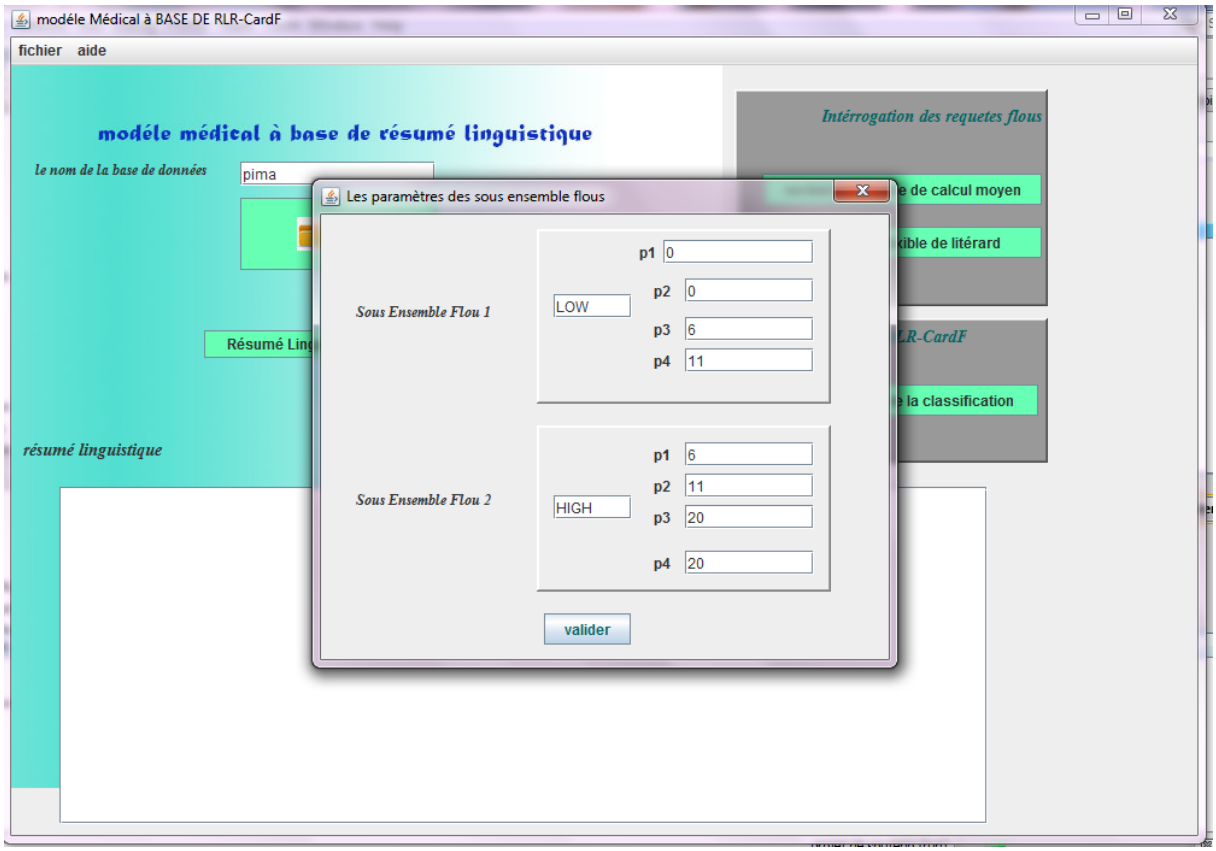
$$F_{UCShape_{LOW}} = 1.0/0 + 1.0/458 + 0.8/501 + 0.5/533 + 0.3/562$$

Nous pouvons dire qu'il ya :

- ✓ 458 patients qui appartiennent totalement à la variable linguistique« faible(low)» de l'attribut « $UCShape$ »(Uniformity of cell shape) .
- ✓ 43 (501-458 = 43) patients qui appartiennent partiellement à la variable linguistique« faible(low)» de l'attribut « $UCShape$ »(Uniformity of cell shape) qui est associé à un degré d'appartenance = 0,8.
- ✓ 32 patients qui appartiennent partiellement à la variable linguistique« faible(low)» de l'attribut « $UCShape$ »(Uniformity of cell shape) qui est associé à un degré d'appartenance = 0,5.
- ✓ 29 patients qui appartiennent partiellement à la variable linguistique« faible (low)» de l'attribut « $UCShape$ »(Uniformity of cell shape) qui est associé à un degré d'appartenance = 0,3.

4.4.3 Aperçue sur l'application

La figure suivante montre l'interface graphique, de notre simulateur, qui donne la possibilité d'entrés les paramètres des variables linguistiques pour chaque attribut, afin de créer un **RLR-CardF** de la base de données **PIMA**



4Figure 4-7 Aperçus sur les entrées des paramètres des variables linguistiques

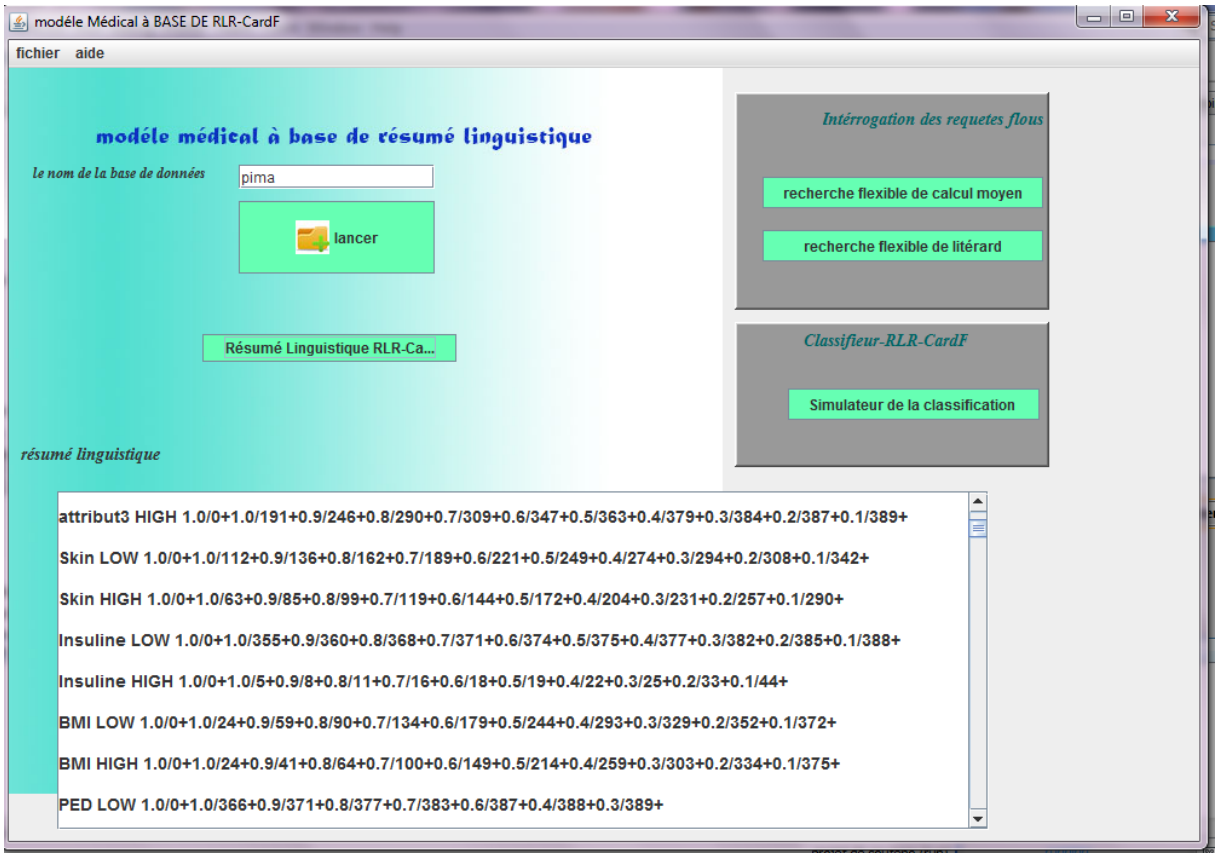


Figure 4-8 Aperçus sur les résumés linguistiques RLR-CardF de la base PIMA

4.5 Approche proposée d'un système d'interrogation médical flexible

Les résumés linguistiques de la base de données fournissent un moyen de réduire considérablement le volume d'entrées. Sachant que l'avantage de tous type des résumés linguistiques est de réduire le temps de réponse, donc, Il est important de trouver un processus de recherche rapide et précis sur nos résumés linguistiques **RLR-CardF**.

Pour un mécanisme d'interrogation efficace, Il faut que la connaissance pertinente puisse être facilement récupérée à partir d'un ensemble de résumés linguistiques. En plus, la réponse demandée doit être précise avec le taux de perte faible.

Par conséquent, il faut créer un système d'interrogation flexible qui répond parfaitement aux besoins de l'utilisateur. Pour cela, nous avons besoin d'identifier le quantificateur flou d'une requête demandée.

Dans les quantificateurs flous introduits par Zadeh, nous trouvons deux type de quantificateurs : absolus (environ 3, au moins 2) ou relatifs (la majorité, environ la moitié).

Les résumés linguistiques, à base de cardinalité floue, de la base de données ''Pima Indiana Diabète'' contiennent des informations quantitatives et des informations qualitatives qui sont interprétables en langage naturel.

En général dans tous types de résumé linguistique, la requête demandée est écrite avec des expressions (proposé par Yager) de langage naturel sous la forme ($Qy'areP$).

Les expressions quantifiées simples sont définies par :

$$"Qy'areP"$$

Exemple :"environ 3 patients sont Glu élevé".

Il existe aussi l'expression de requête complexe qui est définie par :

$$"QRy'areP".$$

Exemple "la plupart des jeunes patients sont Glu faible"

Où Q est un quantificateur linguistique, R et P deux prédicats graduels définis par des ensembles flous et Y une relation de la base de données.

Dans la littérature, nous trouvons plusieurs méthode de calcul ou de validé un quantificateur flou. Nous essayons dans la sous section suivante de présenter quelque méthode de calcul :

4.5.1 Calcul le quantificateur flou par le produit scalaire

La réponse d'une requête peut être calculée, en utilisant les cardinalités floues enregistrés dans nos résumés linguistiques. Une requête simple de la forme $Qy'areP$ ou une requête complexe de la forme $QRy'areP$ qui sont calculées par le produit scalaire [70].

Alors pour calculer le quantificateur flou d'une requête pour un seul prédicat flou P , nous utilisons les informations de résumé F_p .

Donc :

- le produit scalaire de la requête simple est :

$$S_P = \sum_{i=0}^{n} card_i(\lambda_i) \times \lambda_i \Big/ card_i(\lambda_n) \qquad (4.4)$$

Avec $1 > \lambda_1 \geq \ldots \geq \lambda_{n-1} > \lambda_n = 0$ et $n \succ 0$.

Exemple**:** Combien des patients qui ont le Glucose régulier?

Nous Considérons que la requête est simple, nous utilisons le résumé

$F_{regular} = 1.0/0\ 1.0/163\ 0.9/177\ 0.8/191\ 0.7/212\ 0.6/237\ 0.5/262\ 0.4/270\ 0.3/284\ 0.2/295\ 0.1/310$

Alors, le degré scalaire de cette requête est :

$S_{regular} = (163 \times 1) + (177 \times 0.9) + \ldots + (310 \times 0.1) \Big/ 310 \qquad S_{regular} = 0.380$.

Par conséquent, nous déduisons que environ 38% des patients sont Glucose réguliers.

- le produit scalaire d'une requête complexe qui contient deux prédicats flous $P \wedge R$

$$\text{est : } S_{P \wedge R} = \sum_{i=0}^{n} card_i(\lambda_i) \times \lambda_i \Big/ card_i(\lambda_n) \qquad (4.5)$$

Avec $1 > \lambda_1 \geq \ldots \geq \lambda_{n-1} > \lambda_n = 0$ et $n \succ 0$.

Exemple : Combien de jeunes patients qui ont le Glucose réguliers?

Nous calculons le produit scalaire de $F_{regular \wedge young}$.

alors, $F_{regular \wedge young} = 0.52$.

Pour vérifier la précision de la réponse de nos requêtes obtenues par la méthode de calcul de produit scalaire, nous effectuons une autre recherche de ces requêtes sur la base de données numérique médicale. Puis, nous comparons entre ces deux types de recherches .le graphe ci –dessous résume les résultats obtenus :

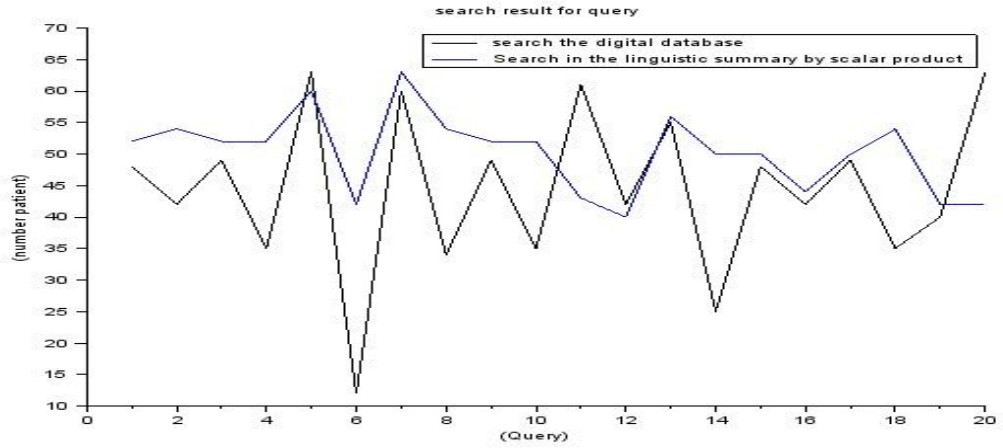


Figure 4-9 Graphe de comparaisons entre deux types de recherches

Dans le graphe4.9, nous remarquons que le traçage de la recherche des requêtes, calculé par le produit scalaire est non identique à celui du système de recherche dans la base de données numérique. Ceci nous amène à conclure que les résultats obtenus par le calcul moyen sont non précis.

4.5.2 Calcul le quantificateur flou par le degré de vérité

Selon le résumé quantifié de Yager, nous validons une requête floue par le degré de vérité de cette dernière.

En général, la requête demandée est écrite avec des expressions (proposé par Yager) de langage naturel sous la forme ($Qy'areP$).

Où le dégréé de vérité de cette requête est :

$$T(Qy'sareS) = \mu_Q\left[\frac{1}{n}\sum_{i=1}^{n}\mu_S(y_i)\right] \qquad (4.6)$$

Dans les résumés linguistiques à base de calcul de la cardinalité floue, le degré de vérité est calculé par expression suivante:

$$T(Q_{t\in R}P(t)) = \sum_{i=0}^{n}(\alpha_i - \alpha_{i+1}) \times \mu_Q\left(i/n\right) \qquad [132] \qquad (4.7)$$

Dans cette approche de recherche, il faut proposer un quantificateur flou d'une requête puis calculé le degré de la vérité de cette dernière. Mais, cette approche reste couteuse puisqu'il y'a un infini de quantificateur flou.

Selon le chercheur L.Liétard [73], il a été prouvé que la protoforme de requête, proposé par Yager , ne présente que les informations quantitatives .Le degré de vérité, de Yager, ne

montre pas suffisamment la relation entre les deux aspects(les informations quantitatives et les informations qualitatives).

Alors, nous proposons, dans notre thèse, un nouveau type de protoforme requête qui est associée à un degré de validité.

4.5.3 L'approche proposée IFlex- RLR-CardF par le degré de la validité

Notre contribution consiste à utiliser le calcul de degré de la validité, proposé par Liétard[72], pour traiter des requêtes floues , dans les résumés linguistiques RLR-CardF.

Nous proposons, pour cela, un nouveau d'algorithme pour la création d'un système d'interrogation médical flexible **IFlex- RLR-CardF**(**I**nterrogation **Flex**ible des **R**ésumés **L**inguistiques à base de calcul de **card**inalité **F**loue) [133] et qui est basé sur le calcul du degré de validité(voir Algorithme4.3).

Nous utilisons l'expression de requête proposée par L.Liétard [72]:

Soit $C^i, C^{i+1}, ..., C^{i+k}$ sont des prédicats flous respectivement définies sur les domaines des attributs $A^i, A^{i+1}, ..., A^{i+k}$ de la relation R .

Les requêtes sont du type: <<typle from R satisfy C^i and $C^{i+1}$and…and C^{i+k} >>

Condition << C^i and C^{i+1} and…and C^{i+k} >>est une contrainte floue exprimée par le résumé.

Le degré de validité d'un tel résumé est donné par:

$$\omega = \max_{\alpha \in [0,1]} \min\left(\alpha, F\left(C_\alpha^i \times ... \times C_\alpha^{i+k}\right)\right) \qquad (4.8)$$

Où F est la fonction définie par :

$$F(E) = \left| E \cap R\left[A^i, ..., A^{i+k}\right]\right| / \left|R\left[A^i, ..., A^{i+1}\right]\right| \qquad (4.9)$$

Et le degré de validité ω garanti le minimum dans les deux termes de quantité et de qualité. Où ω est le plus fort pourcentage de G, tel que G% a au moins n-uplet qui satisfait la requête.

Dans notre travail, nous avons r_{su} qui est l'ensemble de résumés linguistiques à base de calcul de la cardinalité floue.

Nous avons aussi la fonction $F(E)$ qui est exprimé par une intersection entre les prédicats de la requête:

$F(E) = \left| E \cap R\left[A^i, ..., A^{i+k}\right]\right| / \left|R\left[A^i, ..., A^{i+1}\right]\right|$.

Pour calculer la fonction $F(E)$, nous proposons d'effectuer les calculs suivantes :

- effectuons les différents α-cut sur les prédicats C^1 et C^2 de la requête Q, où $\alpha \leq 1 \leq \ldots \leq \delta$.
- calculons $F(E)$ où

$$F\left(C^1, C^2\right)_\alpha = \left|Card C^1 \wedge C^2\right|_\alpha / N \qquad (4.10)$$

avec N: nombre des individus.

L'avantage de notre calcul de F (E), par rapport à la fonction proposé par Liétard, est de calculer la fonction sans le besoin d'accéder aux bases données numériques. Pour cela, nous n'utilisons que les informations des résumés linguistiques.
Dans notre thèse nous proposons un nouveau algorithme de l'interrogation flexible à partir de l'exploitation nos Résumés linguistiques **RLR-CardF**.

Notre contribution est résumée dans l'algorithme4.3 où nous utilisons cette approche pour créer un système d'interrogation médical.

Dans l'algorithme4.3, nous avons $C^1 et\ C^2$ comme prédicats flous de la requête Q. Si δ est le seuil de α-cuts, spécifié par l'utilisateur, il est alors nécessaire d'évaluer uniquement l'α-cuts tel que $\alpha \geq \delta$.

Algorithme4.3 : IFlex- RLR-CardF qui est basé sur **calcul le degré de validité d'une requête complexe**

Input Q is query
Intput δ threshold
Input $C^1 and\ C^2$ fuzzy predicate.
Input r_{sum} summaries linguistic
Output validity degree
For α:=1**down to** δ **do**
Compute $F\left(C^1, C^2\right)_\alpha = \left\|Card C^1 \wedge C^2\right\|_\alpha / N$
Compute validity degree
End for

Nous présentons par la suite un exemple qui s'explique le processus de notre algorithme d'interrogation qui cherche des réponses précis des requêtes floues :

Soit A1 « Glucose» **Glu** est un attribut.et A2 «**Age**» est un deuxième attribut

Selon le processus de FUZZIFICATION, l'attribut **Glu** est partitionné en deux sous ensemble flou (faible, élevé).l'attribut **Age** est partitionné en deux sous ensembles flous (jeune, âgé).

Selon, le RLR-CardF est représenté par : $F_{faible \wedge jeune}$.

$F_{faible \wedge jeune}$ =1.0/0+0.9/4+0.8/16+0.7/42+0.6/64+0.5/102+0.4/129+0.3/153+0.2/184+0.1/232

Dans cet exemple, nous voulons chercher la réponse d'une requête floue de la forme :<<**les patients qui satisfont Glu faible et Age jeune >>.**

Selon l'alogithme4.3 :

- Nous proposons un seuil δ qui est égale à0.2.
- Nous effectuons diffèrent α-cut sur les prédicats C^1 et C^2 de la requête Q, où $\alpha \leq 1 \leq ... \leq \delta$. et C^1 est une variable linguistique faible-Glu et C^2 est une variable linguistique (jeune-Age).

Nous obtenons les résultats dans le tableaux4.3 et le tableau 4.4 :

Tableau 4-3Tableau de α -cut de C_α^1 and C_α^2

α	C_α^1	C_α^2
1	$[0,42]$	$[0,38]$
0.8	$[0,44]$	$[0,43]$
0.6	$[0,50]$	$[0,47]$
0.4	$[0,54]$	$[0,52]$
0.2	$[0,52]$	$[0,57]$

Tableau 4-4 Tableau de la fonction Graduel

α	$F\left(\mathrm{C}^1,\mathrm{C}^2\right)_\alpha = \left\lvert Card C^1 \wedge C^2 \right\rvert_\alpha / \mathrm{N}$
1	0
0.8	0.18
0.6	0.49
0.4	0.72
0.2	0.89

$\omega = max\left(min(1,0)\ min(0.8,0.18) min(0.6,0.49), min(0.4,0.72)\ min(0.2,0.89)\right)$

ω =0,49

Alors nous pouvons dire que **''Environ 49% des patients ont Glu regular et age young''** avec le temps de réponse =1 milliseconde.

Selon l'algorithme 4.3, nous avons effectué la recherche des réponses sur plusieurs requêtes floues où nous obtenons, à la fin, une base de connaissances qui est interprétable en langage naturel:

Tableau 4-5la base de connaissances

Requête	Ω	Temps de réponse
Satisfy glu normal et age young	0.49 (49%)	1 milliseconde
Satisfy bp hight et insuline low	0.59 (59%)	1 milliseconde
Satisfy glu normal et insuline low	0.32 (32%)	2milliseconde
Satisfy bmi obésity et age old	0.10 (10%)	1 milliseconde

Selon les tableaux 4.5, nous construisons une base de connaissances qui peut d'aider le médecin de visualiser touts les cas des patients, par exemple : les symptômes des patients diabétiques.
Les avantages de notre approche de l'interrogation sont:

- Minimiser le temps de recherche des informations où la recherche d'une requête dans notre système est effectuée pendant 1milliseconde et la recherche d'une requête dans une base de données numérique est effectué pendant1 seconde 50 millisecondes.
- La base de connaissance contient toutes les informations quantitatives et qualitatives utiles pour le médecin.
- Les résultats sont interprétés en langage naturel et sont faciles à lire.
- Le degré de validité offre la précision de la requête, dans le résume linguistique à base de calcul de la cardinalité floue RLR-CardF. Il est utile pour l'utilisateur (médecin) pour prendre la meilleure décision.

Le graphe ci-dessous représente une comparaison entre la recherche de la réponse d'une requête flexible dans la base de données numérique et la recherche de la réponse de la même requête, à base de notre approche proposé **IFlex- RLR-CardF, dans** nos résumés linguistiques.

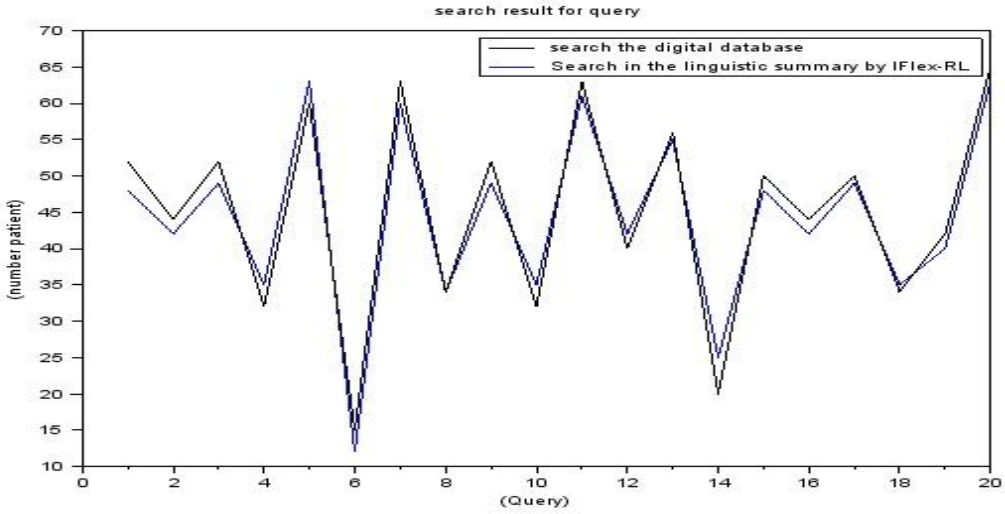


Figure 4-10Résultat de la recherche d'une requête par les méthodes de recherches

Dans le graphe4.10, nous remarquons que le traçage du système d'interrogation, développé dans cette thèse est presque identique à celui du système d'interrogation de la base de données numérique. Ceci nous amène à conclure que nos résultats sont précis.

En conséquence, nous pouvons dire que notre algorithme 4.3 est rapide et précis et que notre système d'interrogation médical, proposé dans notre thèse, est flexible.

4.5.4 Aperçue sur l'application

La figure suivante montre l'interface graphique de l'exécution de notre simulateur afin de créer un **IFlex**- **RLR-CardF** de la base de données **médical PIMA**

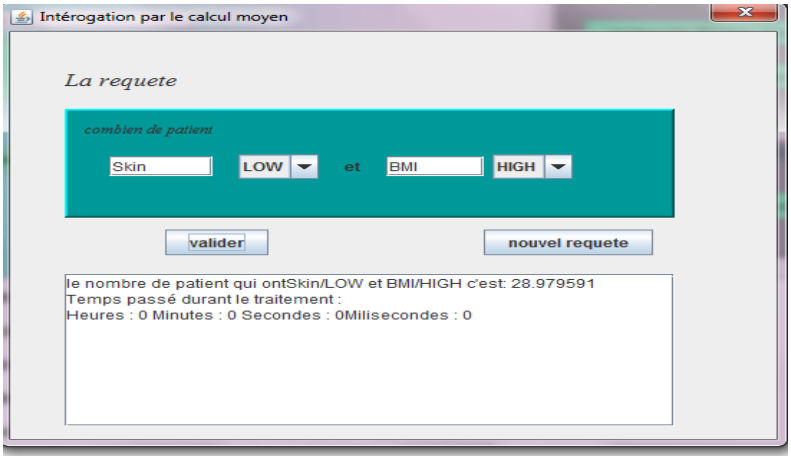


Figure 4-11 Aperçus sur la réponse d'une requête floue

4.6 Approche proposée d'un classifieur médical « Classifieur- RLR-CardF»:

Dans le domaine de résumé linguistique, les chercheurs ont souligné l'intérêt des résumés linguistiques dans les systèmes de déduction. Mais, ils ne l'ont pas intégrer réellement dans le système de diagnostic médical.

Dans des travaux similaires, les chercheurs ont effectué une classification supervisée sur les bases de données médicales, en utilisant les méthodes de classification traditionnelle comme SVM, réseaux de neurones, K-NN, etc. [134][135]. Mais les résultats obtenus n'ont pas été compréhensible par les médecins. En plus, ces méthodes ne peuvent pas souvent traiter une grande masse de données où le temps d'exécution sera lent.

Notre contribution est d'utilisé les résumés linguistiques médicaux **RLR-CardF** afin de crée un classifieur fiable et rapide (L'état de patient).

4.6.1 Schéma général de l'approche proposé « Classifieur- RLR-CardF »

Nous présentons dans le graphe suivant, le schéma de notre approche proposée de la classification supervisé de données médicales. Cette approche est basée sur le calcul de la similarité entre les résumés linguistiques flous.

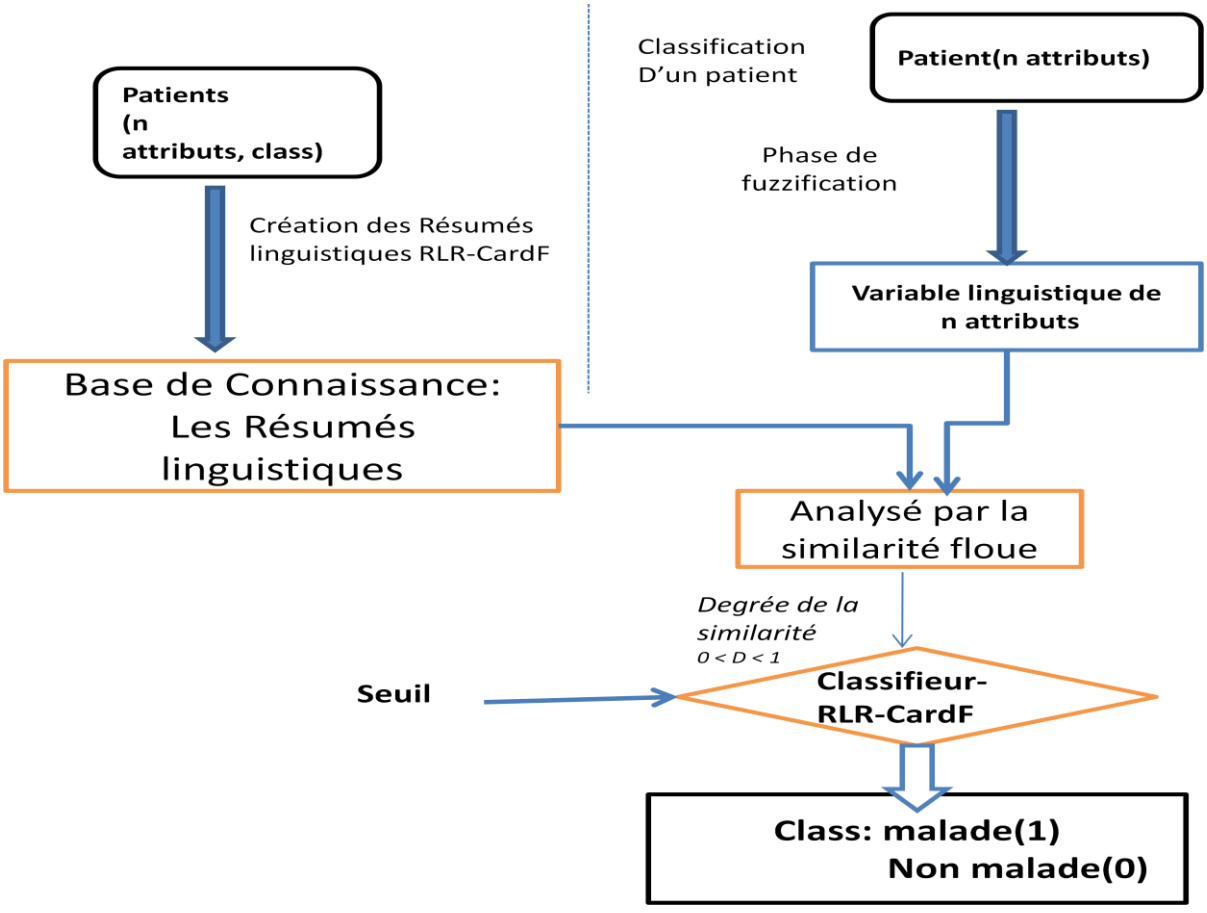


Figure 4-12 Schéma général de notre approche de la classification supervisé

Cette figure(figure4.12) présente le schéma générale de notre Approche de la classification supervisée **Classifieur- RLR-Card**[136] , à base de résumé linguistique. Ce modèle est composé en deux phases :

- La première phase, c'est la construction des résumés linguistiques **RLR-CardF**. Elle est considérée comme une phase d'apprentissage de notre système de la classification. Son principe est de créer des résumés linguistiques à partir d'une base de données numérique où nous construisons, à l'aide de l'algorithme4.2, les résumés linguistiques **RLR-CardF**. Nous choisissons ce type de résumé parce qu'il contient toutes les informations quantitatives et les informations qualificatives essentielles pour une prise de décision efficace. Donc, ces informations sont ajoutées pour construire une base de connaissances riche.
- Dans la deuxième phase, Il est possible de faire une classification supervisée à l'aide des informations obtenues à partir des résumés linguistiques médicaux. Afin de prédire l'état de santé d'un patient, nous utilisons pour cela notre base de connaissances et le calcul de la similarité, proposé par Wilbek où le degré de la similarité appartient à l'intervalle [0,1]. Nous obtenons à la fin, l'état de santé du patient (malade ou non) .En suite, nous enrichirons la base de connaissance par les informations obtenues (mise à jour).

4.6.2 Processus de développement, le Classifieur- RLR-CardF

Dans les travaux antérieurs, l'utilisation du degré de similarité avait pour but de réduire l'ensemble des résumés linguistiques, par la comparaison entre les résumés linguistiques et par la suppression des résumés qui ont le même poids [121].

Même, il existe des autres travaux où ils ont exploité les résumés linguistiques dans les systèmes de la classification classique (SVM,RN,….)mais les résultats obtenus sont faibles [130].

Par contre, le principe de notre approche de la classification est d'exploiter les résumé linguistiques qui sont considéré comme un support de données, ainsi que le calcul de similarité entre les résumés linguistiques, afin déduire l'état de patient.

Prenons comme hypothèse : un patient est considéré comme une entrée avec n attributs $a_1,...,a_n$ où $a_1,...,a_n \in R$. Puis, nous faussons la fuzzification qui converti chaque valeur numérique en une ou plusieurs variables linguistiques, avec les degrés d'appartenances. Ensuite, effectuons l'algorithme 4.2, afin de construire un ensemble de résumés linguistiques.

Chaque résumé linguistique est sous forme d'un ensemble des cardinalités floues qui est défini par : $F_{a_1}, F_{a_2}, .., F_{a_1 \wedge a_n}, ..., F_{a_{n-1} \wedge a_n}$.

L'idée principale est de créer une matrice de la similarité, qui contient les degrés des similarités entre les variables linguistiques des attributs et l'intersection les mêmes variables linguistiques avec la variable linguistique « malade», de l'attribut : classe. Notons que l'attribut (classe) représente l'état de santé du patient qui prend une variable linguistique (malade, non malade).

Pour créer une matrice de similarité, il faut calculer:

$Sim(\mathrm{L}S_1, \mathrm{LS}_2)$ .Où : $LS_1 = Qy'areP$ et $LS_2 = Qsicky'areP$.

Dans notre thèse, nous affectons le quantificateur flou Q ''**plus** ''.

Donc, nous avons : $yQ(x) = x$.

En choisissons le résumé linguistique à base de calcul de cardinalité floue, le degré de la vérité est calculé par la méthode proposée dans [132] :

$$T(Q_{t \in R}P(\mathrm{t})) = \sum_{i=0}^{n} (\alpha_i - \alpha_{i+1}) \times \mu_Q\left(i/n\right) \qquad (4.11)$$

Où α_i est le degré d'appartenance à l'attribut qui est représenté par λ_i dans notre résumé linguistique F_A .

Sachant que le degré d'appartenance est incrémenté dans intervalle [1, 0] et $\lambda_0 = 1$.

Donc, nous avons :

$$\mu_Q(i/n) = {card_i(\lambda_i) - card_{i-1}(\lambda_{i-1})} \Big/ {card_i(\lambda_n)} \qquad (4.12)$$

Selon[121], la similitude du résumé linguistique est représenté par l'équation suivante :

$$sim(\mathrm{LS}_1, LS_2) = \min(sim(\mathrm{P}_1, \mathrm{P}_2), sim(\mathrm{Q}_1, \mathrm{Q}_2), sim(\mathrm{R}_1, \mathrm{R}_2), sim(\mathrm{T}_1, \mathrm{T}_2)) \qquad (4.13)$$

Cependant, dans le travail d'Almeida [137], il a été prouvé que la dissemblance est juste la différence entre les valeurs de T, qui représente le degré de vérité d'un résumé linguistique.

$$\mathrm{d}(\mathrm{LS}_1, LS_2) = \left|T_1 - T_2\right| \qquad (4.14)$$

Sachant que :

$$sim(\mathrm{LS}_1, LS_2) = 1 - d(\mathrm{LS}_1, LS_2) \qquad (4.15)$$

Donc, selon les équations 4.14 et équation 4.15 nous obtenons:

$$sim(\mathrm{LS}_1, LS_2) = 1 - \left| T_1 - T_2 \right| \qquad (4.16)$$

Notre contribution est utilisé la formule de degré de vérité de RL-CardF pour déduire $sim(\mathrm{LS}_1, LS_2)$

Nous effectuons ces calculs sur m variables linguistiques pour trouver une matrice de la similarité de dimension $m \times 1$ où nous calculons la similarité juste pour les variables linguistiques qui représentent un facteur de risque.

- Par exemple : le facteur de risque de l'attribut Glucose dans un sous-ensemble flou est représenté par la variable linguistique (élevé).

La matrice de la similarité est : $\begin{bmatrix} sim(\mathrm{L}\,S_{P_1}, LS_{P_1 \wedge P_{malade}}) \\ sim(\mathrm{L}\,S_{P_2}, LS_{P_2 \wedge P_{malade}}) \\ \vdots \\ sim(\mathrm{L}\,S_{P_m}, LS_{P_m \wedge P_{malade}}) \end{bmatrix}$.

A fin de raffiner notre matrice de similarité et de trouver l'état du patient à classer, nous utilisons l'une des méthodes d'agrégations qui est l'intégrale de Sugeno, prise entre chaque degré de similarité d'une variable linguistique et le α_cut où $\alpha \in [1,0]$ .L'intégrale est définie par :

$$C = \max(\min(\alpha, f(\alpha))) \qquad (4.17)$$

Dans notre cas, f(α) est une fonction de similarité pour chaque variable linguistique qui représente le facteur de risque de nos résumés linguistiques et α_coupe : $\alpha \in [1,0]$.

$$C = \max(\min(\alpha_i, sim(LS_1, LS_2))) \qquad (4.18)$$

Notre contribution vise à utiliser les résumés linguistiques **RLR-CardF** médicaux dans un système de la classification supervisée, afin de déduire l'état de santé des patients. Il est possible de trouver la classe du patient (malade ou non malade), en utilisant l'intégrale Sugeno entre les degrés d'appartenances des variables linguistiques du patient et le degré de la similarité des variables linguistiques. Ceci est détaillé dans l'algorithme4.4 suivant :

Algorithme 4.4 : classification supervisé à base des résumés linguistiques, Classifieur- RLR-CardF

Input A is attribute, N number of attributes

Input r_{sum} are linguistic summaries
Output the validity degree C
Process:
L= 0;
For i←1 to N do
$(S_i, \mu_i) = fuzzification(A_i)$ **;**
L←L+1;
End for;
For i ← 0 **to L do**
$T_1 \leftarrow$ degree of truth(LS_1) **;**
$T_2 \leftarrow$ degree of truth(LS_2) **;**
$sim(\mathrm{LS}_1, LS_2) \leftarrow 1 - |T_1 - T_2|$ **;**
End for ;
$C \leftarrow \max(\min(\mu_i, sim(LS_1, LS_2)))$ **;**
Return C

Nous expliquons l'exécution de notre algorithme de classification supervisé, à travers cet exemple :

Choisissons la base donnée médicale ''PIMA Indiana diabète ''.

Le seuil pour déduire la classe est 0.8.

Nous avons un Patient avec 8 attributs :

	NbGross	Glu	BP	SKIN	Insulin	BMI	PED	Age
Patient	2	179	70	45	543	30.5	0.15	53

Il est bien connu que: $sim(\mathrm{LS}_1, LS_2) = 1 - |T_1 - T_2|$

Suite aux hypothèses énoncées et a l'exécution de l'algorithme4.4, nous aurons les tableaux suivant :

Tableau 4-6 Tableau de la Fuzzification des attributs d'un patient

	NbGross	Glu	BP	Skin	Insulin	BMI	PED	Age
SE	Regular High	Regular High	Regular High	Regular High	Regular High	Regular High	Regular High	jeune âgé
μ	0 1	1 0	0.625 0.375	0 1	0 1	1 0	1 0	0.375 0.625

Le tableau suivant indique le calcul de T1 et T2 pour chaque sous-ensemble flou(SE).

Tableau 4-7Table T1, T2

	NbGross	Glu	BP	Skin	Insulin	BMI	PED	Age
SE	**Regular High**	**Regular High**	**Regular High**	**Regular High**	**Regular High**	**Regular High**	**Regular High**	**jeune âgé**
T1	0,9 0.8	0.8 0.6	0.6 0.5	0.8 0.6	0.9 0.4	0.55 0.69	0.98 0.35	0.923 0.0.41
T2	0.6 0.5	0.4 0.5	0.5 0.5	0.5 0.5	0.7 0.3	0.365 0.597	0.76 0.23	0.58 0.41

Où T1 est le degré de vérité de chaque sous-ensemble flou et T2 le degré de vérité de l'intersection de chaque sous-ensemble flou avec la variable linguistique (malade).

Calculons la matrice de similarité entre LS1 et LS2, nous obtenons le tableau4.8. Sachant que

$$sim(\mathrm{LS}_1, LS_2) = 1 - \left| T_1 - T_2 \right|$$

Le tableau4.8 affiche le tableau de la similarité entre les variables linguistiques.

Tableau 4-8 Tableau de la similarité

	NbGross	Glu	BP	Skin	Insulin	BMI	PED	Age
SE	**High**	**High**	**High**	**High**	**High**	**High**	**High**	**âgé**
α_i	0	0	0.375	1	1	0	0	0.625
Sim	0.7	0.9	1.0	0.9	0.9	0.9	0.87	0.98

Nous déduisons la classe avec le calcul de l'intégrale sugeno :

On a $C = \max(\min(\alpha_i, sim(LS_1, LS_2)))$

En utilisant les paramètres du tableau5, nous obtenons :

$C = \max(0.0, 0.9, 0.37, 0.9, 0.9, 0.0, 0.0, 0.62)$

$C = 0.9$

La classe $C \geq 0.8$ donc, ce patient est classé diabétique.

Après vérification dans la base de données numérique PIMA, l'attribut class = 1 est trouvé, ce qui signifie que ce résultat est vrai positif (TP).

Citons un autre exemple :

	NbGross	Glu	BP	SKIN	Insulin	BMI	PED	Age
Patient	2	99	60	17	160	36.6	0.453	21

Après l'exécution de l'algorithme4, on a trouvé $C = 0.0625$

La classe $C \leq 0.8$ donc le patient est classé non diabétique.

Après vérification dans la base de données PIMA, l'attribut class = 0 est trouvé, ce qui signifie que ce résultat est vrai négatif (TN).

4.6.3 Aperçue sur l'application

La figure suivante montre un schéma de l'exécution afin de créer un **Classifieur- RLR-CardF** de la base de données **PIMA**

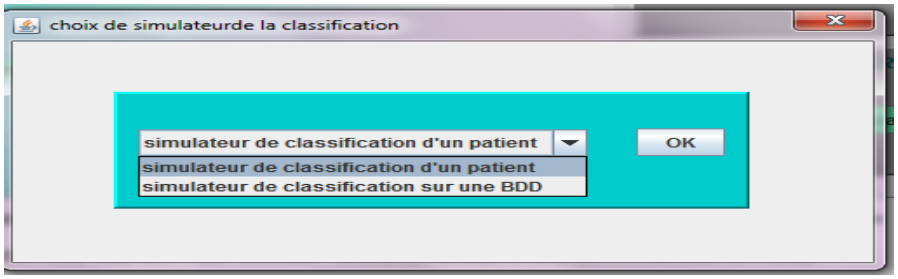


Figure 4-13Fenêtre de choix de simulateur de la classification

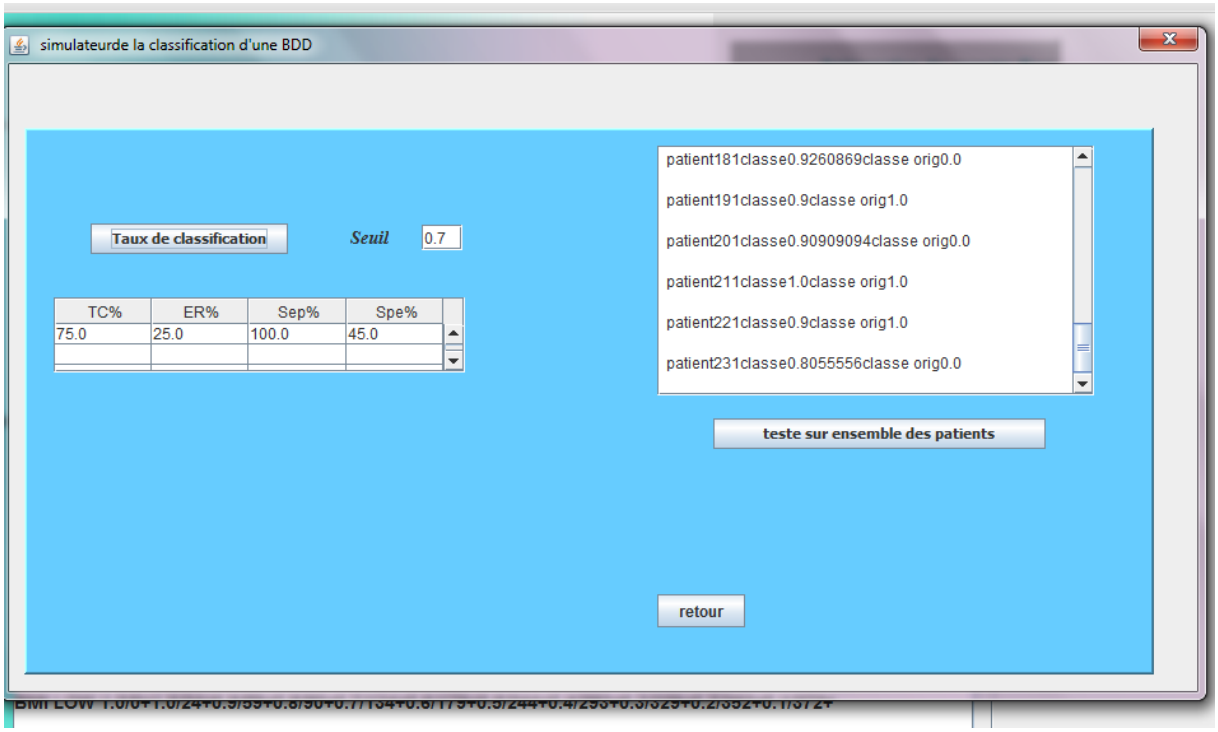


Figure 4-14Fenêtre de l'exécution du simulateur de la base de test

4.7 Résultats et discussions

Dans cette section, nous allons présenter l'évaluation des performances de notre modèle de classification supervisé médical

Différents jeux de données sont utilisés pour évaluer les performances de cette nouvelle approche de la classification **classifieur-RLR-CardF**. Ces jeux de données ont été pris d'UCI Machine Learning Repository. (Pima, WDBC, et mamographie).

Dans notre système de classification, nous découpons chaque base de données en deux parties où 4/5 des instances sont utilisées dans la phase d'apprentissage, est1/5 dans la phase de test et la validation.

Les performances du système de notre classifieur flou **Classifieur-RLR-CardF** est évaluée en utilisant les paramètres suivants: taux de classification correct, taux d'erreur, sensibilité et spécificité.

Où : CC = VP+VN/ (VP+VP+FP+FN)*100 est le taux de classification correct.

Error rate = FP+FN/(VP+VP+FP+FN)*100 est le taux d'erreur.

Se: Sensibilité = VP/(VP+FN)*100 est le taux de vrai positif,

Sp: Specificité = VN/(VN+FP)*100 est le taux de vrai négatif.

Sachant que VP, VN, FP et FN désignent respectivement : vrais positifs, vrais négatifs, faux positifs et de faux négatifs.

La sensibilité et la spécificité fournissent une estimation de performance d'un classifieur, ils déduisent la capacité de prédiction d'une méthode de classification.

Dans notre étude, nous s'intéressons particulièrement sur la capacité prédictive pour les classes minoritaires .Donc, un bon classifieur réalise une bonne connaissance des données positives.

Le Tableau 4.9 présente les résultats de paramètre de performance obtenus sur trois bases de données médicales.

Tableau 4-9Tableau de performance de Classifieur-RLR-CardF

	PIMA	**WBCD**	**mammographie**
Classification rate	83%	96%	80%
Error rate	17%	4%	20%
Sensibility	100%	100%	100%
Specificity	63%	94%	58%

Dans ce modèle, nous avons validé notre algorithme4.4 sur plusieurs base de données .Les performances de la méthode proposée ont été testées par le taux de classification, la sensibilité et la spécificité du classifieur à partir de résumé linguistique.

L'analyse des résultats décrit dans le Tableau 4.9 montre que notre Classifieur-RLR-CardF a donné de très bons résultats et a atteint un taux de classification correcte de 96%. Ce résultat est enregistré pour les jeux de données : PIMA, WBCD, MAMOGRAPHY.

La figure suivante résume le taux de classification, taux d'erreur, sensibilité et spécificité pour chaque jeu de données où nous pouvons faire une comparaison sur la performance de notre classifieur.

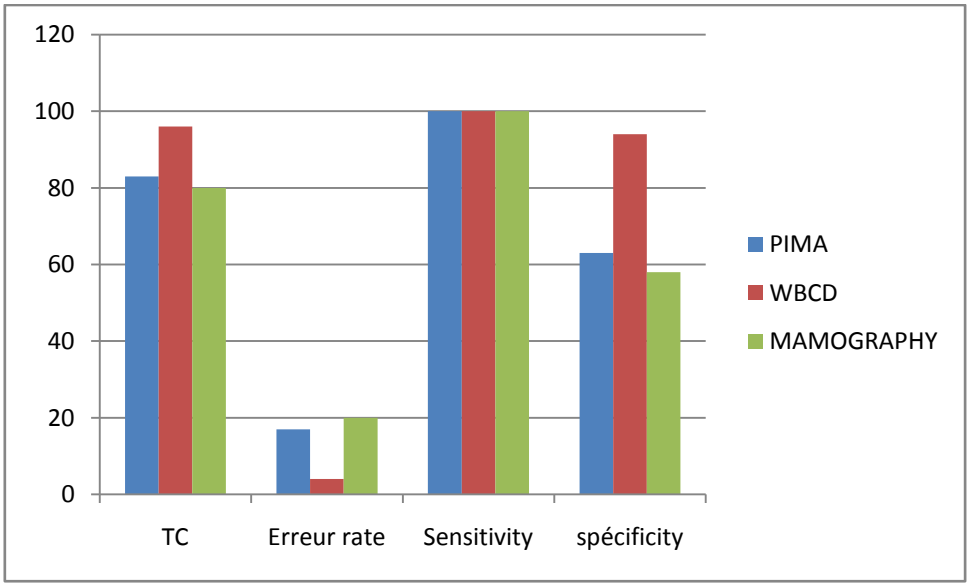


Figure 4-15Comparaisons la performance de notre classifieur sur des bases de données médicales

Selon le tableau 4.9et la figure4.15, nous remarquons que le meilleur de taux de classification obtenu est 96% de la base de données **WBCD** .Le taux de classification est élevé, ce qui signifie que notre système fournit des résultats efficaces sur la base de données **WBCD**. Alors, le taux de classification moyen obtenu par notre algorithme est de 83,75%. Aussi, nous remarquons que taux d'erreur est faible où la moyenne de taux d'erreur est 16.25%.

Aussi, le point fort de notre système c'est le taux de sensibilité qui est arrivé jusque 100% dans la plupart des bases de données. Par conséquent, nous avons obtenu les meilleures performances de classification. Nous pouvons dire que notre classificateur a une bonne reconnaissance des classes minoritaires. Nous déduisons que les performances de classification (SE, SP, TC) sont grandes par contre le taux erreur est petit. Aussi, la sensibilité obtenue par le notre classifieur est très grande ce, qui veut dire que notre **Classifieur-RLR-CardF** a réalisé une bonne reconnaissance des données positives.

Notre algorithme de classification a besoin d'un seuil pour valider une classe de patient, sachant que le seuil de validation varie selon la base de données. Alors, il faut trouver le meilleur seuil de validation pour obtenir un meilleur taux de classification. Nous avons effectuée plusieurs tests avant de décider le choix de meilleur seuil pour un meilleur taux de classification.

Le taux de classification correct est calculé pour différents seuil de validation. Nous exécutons l'algorithme4.4 neuf fois où le seuil choisi appartient à l'intervalle [0,1]. Par la suite, nous calculons le taux de classification correct et nous sauvegardons le meilleur, le mauvais et le moyen taux.

Tableau 4-10 les résultats obtenus par notre approche de différent seuil de validation.

Seuil	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Pima %	52	54	54	54	54	62	75	75	83
WBCD%	76	76	88	88	88	92	96	88	65
Mamographie%	53	53	53	53	53	65	73	80	69

Le Tableau 4.10- montre le taux de classification correcte obtenu par notre modèle pour chaque base de données. Le tableau montre également le meilleur, le moyen et le mauvais taux de classification correcte produit pour chaque seuil de validation.

Nous remarquons que chaque base de données a un seuil qui diffère de l'autre et qui donne un meilleur résultat de classification.
Le graphe ci-dessous montre le taux de classification de la base de données PIMA de différents seuils de validation

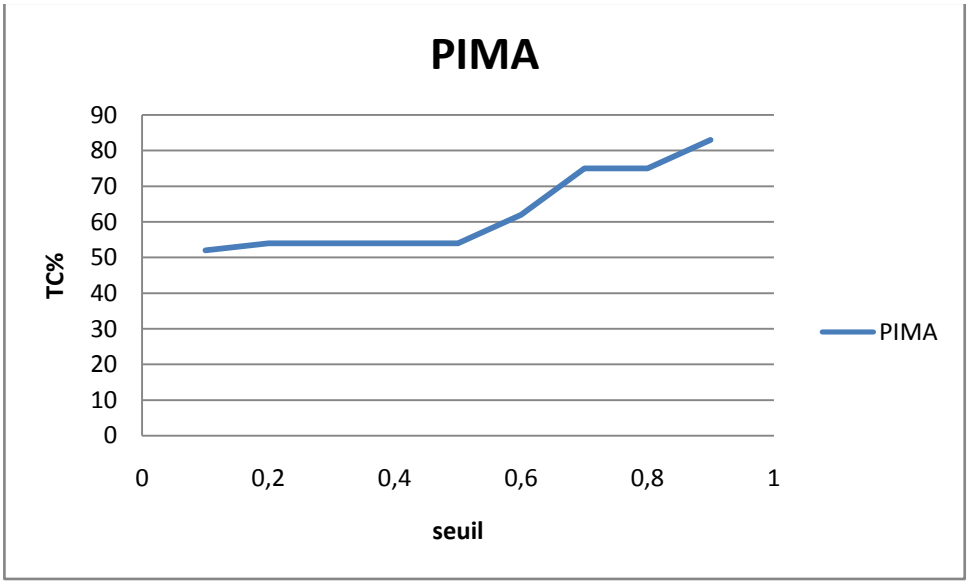


Figure 4-16Le taux de classification Classifieur-RLR-CardF de la base de données PIMA des différents seuils de validation

D'après le traçage de graphe (figure4.16), nous déduisons que le meilleur seuil de la base de données PIMA qui donne un meilleur taux de classification est 0.9.

Le graphe ci-dessous montre le taux de classification de la base de données WBCD de différents seuils de validation

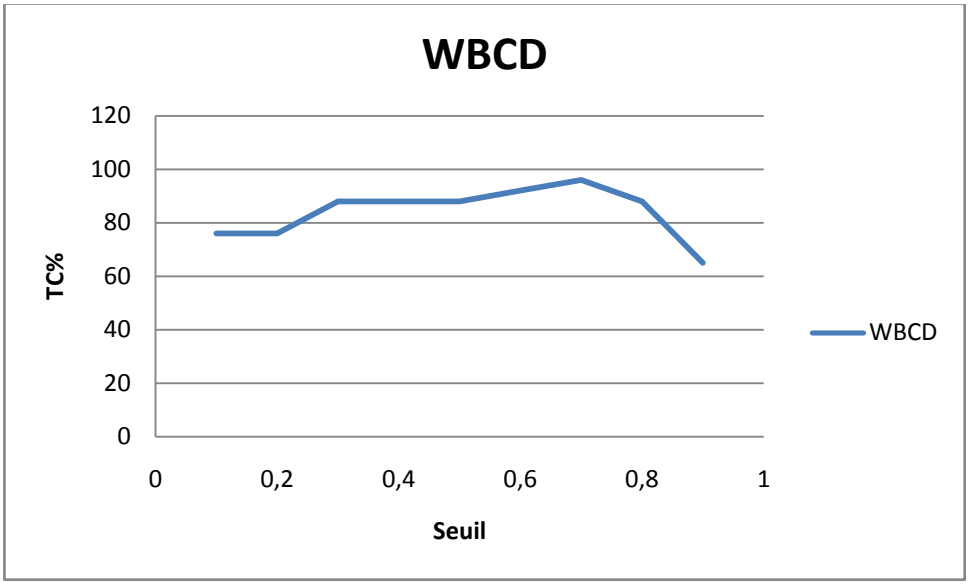


Figure 4-17 taux de classification Classifieur-RLR-CardF de la base de données WBCD des différents seuils de validation

D'après le traçage de graphe (figure4.17), nous déduisons que le meilleur seuil de la base de données WBCD qui donne un meilleur taux de classification est 0.7.

Le graphe ci-dessous montre le taux de classification de la base de données Mamographie, des différents seuils de validation.

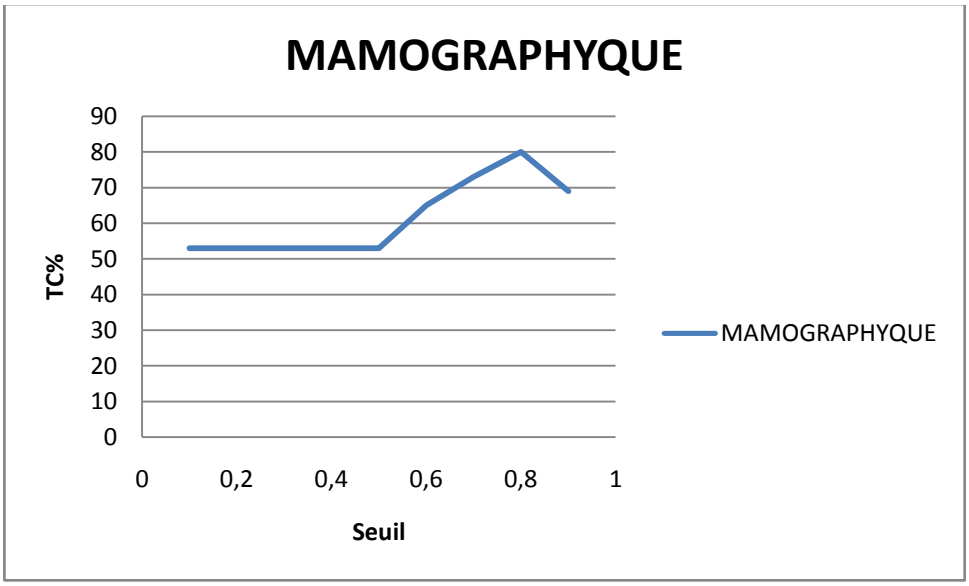


Figure 4-18 taux de classification Classifieur-RLR-CardF de la base de données Mamographie de différents seuils de validation

D'après le traçage de graphe (Figure 4.18), nous déduisons que le meilleur seuil de la base de données MAMOGRAPHY qui donne un meilleur taux de classification est 0.8.

Nous présentons par la suite un graphe qui compare le taux de classification de plusieurs bases de données et qui indique le choix de meilleur seuil :

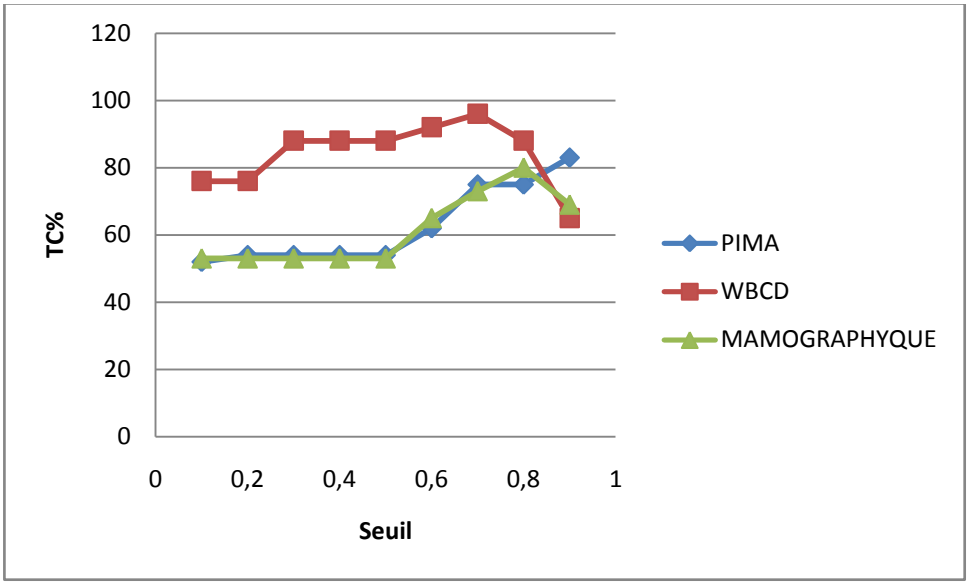


Figure 4-19 Comparaisons le taux de classification de notre modèle Classifieur-RLR-CardF effectué sur des bases de données médicales

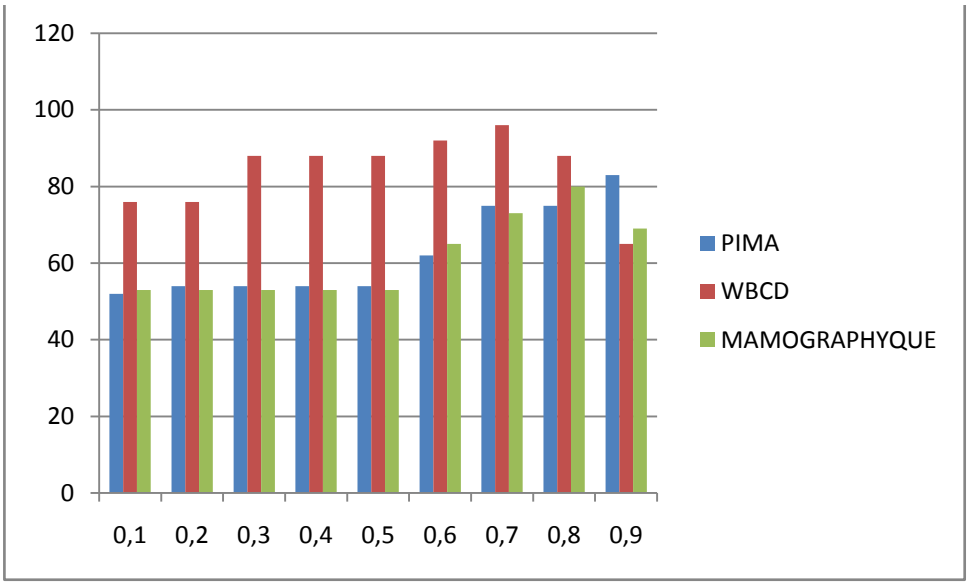


Figure 4-20 Histogramme de taux de classification de notre modèle Classifieur-RLR-CardF sur les bases de données

Selon (figure 4.19) (figure4.20), nous choisissons le meilleur seuil après plusieurs exécutions dans différentes bases de données. Nous observons dans notre modèle que nous trouvons un très bon résultat dans un seuil spécifique.

En conclure, nous pouvons dire que notre système de classification est efficace, facile et interprétable puisqu'il repose sur la théorie de la logique floue. Nous avons développé un classifieur des résumés linguistiques **Classifieur-RLR-CardF** au lieu de classifier sur des bases de données numériques.

4.8 Comparaison de l'approche proposée avec les travaux de la littérature

Dans cette section, nous avons comparé les précisions de taux de la classification de notre modèle, qui à basé sur les résumés linguistiques **RLR-CardF**, avec d'autres méthodes de classifications qui sont appliquées à partir des bases de données numériques :

Les résultats obtenus pour les jeux de données (PIMA, WBCD, Mamographie) de notre modèle, comparés avec ceux des autres chercheurs dans la littérature, sont présentés dans le tableau suivant:

Tableau 4-11Tableau comparatif de notre travail avec les travaux de littératures

Les Bases de Données	La littérature	TC %
PIMA	GSVM[138]	74.15
	WLTSVM[139]	76.78 ±0.35
	GA-Fuzzy classifier [140]	71.49
	FCM.ANFIS[141]	83.85
	C4.5[142]	74.48
	Notre Approche Classifieur- RLR-CardF[136]	83
WBCD	S- AIRS[143]	96.91
	WLTSVM[139]	96.30±0.31
	Bayesian classifiers [143]	92.80
	HMM.Fuzzy[144]	97.07
	AS-NN [145]	95.60
	Notre Approche Classifieur- RLR-CardF[136]	96
Mamographie	MRMR[146]	83.17
	MRF[147]	78,77
	CIFE [148]	62,98
	FISHER[149]	79,53
	IMI[150]	82,98
	Notre Approche Classifieur- RLR-CardF[136]	80

Nous pouvons remarquer à travers ce tableau (tableau 4.11) que le taux de la classification obtenu par notre modèle est meilleur où équivalant aux autres classifieurs développés à partir des bases de données numériques médicales.

Donc, Nous résumons que notre modèle de diagnostic médical traite le problème de volumes de la base de données médical où le résumé linguistique est basé sur le calcul de la cardinalité floue. Ces résumés **RLR-CardF** nous ont aidés à obtenir des résultats efficaces et rapides.

En plus l'utilisation de la notion de la logique floue facilite la tache à l'expert qui lui permet de mieux interpréter les résultats.

Pour conclure, notre modèle de diagnostique médical est précis et interprétable en langage naturel. Il combine entre la création des résumés sémantiques des données, le système de l'interrogation des requêtes floues et le système de classification supervisée. Les résultats obtenus dans cette thèse sont très prometteurs et peuvent être utilisés pour aider le médecin dans ces diagnostics.

4.9 Conclusion

Notre modèle médical, proposé dans notre thèse, est adaptable pour tous volumes de données. Il combine entre la création des résumés linguistiques **RLR-CardF,** le développement d'une nouvelle approche pour le système de requête flexible et la proposition d'un nouveau classifieur supervisé flou. Nous obtenons une base de connaissances qui sont interprétables en langage naturel. Notre approche proposée, **IFlex-RLR-CardF**, effectue une recherche de la requête floue sur notre base de connaissances (ensemble des résumés linguistiques **RLR-CardF)** avec une rapidité et une précision. Ce qui nous permet de créer un système d'interrogation médical flexible.

Notre approche proposé « **Classifieur- RLR-CardF»** d'un système de classification supervisé médical est un classifieur fiable et robuste car nous jugeons la fiabilité de notre approche à travers le taux de classification qui est élevé. Nous effectuons le test sur 3 bases de données médicales. Nous remarquons aussi que le taux de sensibilité est aussi élevé dans plusieurs bases de données où nous trouvons que le taux de sensibilité est arrive jusque 100% c.à.d. FP (faux positive) est 0% ou proche 0. Ce taux signifie que notre classifieur traite parfaitement une connaissance des données positives. Ceci est important pour le médecin afin d'éviter de faire des faux diagnostics.

Conclusion général

Les travaux présentés dans cette thèse ont abordé le problème de l'extraction et la déduction des connaissances à partir d'une grande masse de données médicales, nous avons traité en particulier le problème de la taille volumineuse de données médicales. Nous avons utilisé les informations des résumés linguistiques, à base de calcul de cardinalité floue, afin de construire un classifieur explicite et performant.

Nos principaux objectifs, ont été d'apporter des contributions sur construction d'un résumé linguistique sur les bases de données médicales, la création d'un système d'interrogation médical flexible et de réaliser un classifieur supervisé basé sur ces résumés linguistiques.

Le modèle que nous avons proposé, pour atteindre ces objectifs, a été testé sur des bases de données réelles, issues de plusieurs bases de données médicales existantes dans la littérature.

Dans la première partie, Nous avons proposé d'améliorer la méthode de résumé linguistique à base de calcul de cardinalité floue, proposé par Pard et Dubois, pour traiter le problème de grand volume de données médicales. Nous avons proposé un algorithme de résumé linguistique, à base de calcul de cardinalité floue réduit (**RLR-CardF),** qui est interprétable et proche du langage humain (médecin).

Les résumés linguistiques **RLR-CardF** condensent toutes les informations avec leur poids et leurs degrés de satisfactions .Ils utilisent la notion de la logique floue qui offre aux résumés une meilleure représentation des connaissances, qui sont proche du langage naturel.

Dans la deuxième partie de notre modèle est utilisé les résumés linguistiques pour construire un système d'interrogation médical rapide et flexible, et qui répond aux besoins spécifiques de l'utilisateur.

L'exploitation de résumé linguistique, à base de calcul de cardinalité floue, donne la possibilité de faire une recherche par des requêtes sémantiquement proches au langage naturel.

Notre contribution concerne le calcul de degré de la validité d'une requête, proposé par L.Liétard, pour valider une requête complexe et de trouver le quantificateur flou d'une requête **IFlex- RL** (**I**nterrogation **Flex**ible des **R**ésumés **L**inguistiques). Nous constatons que les résultats obtenus, à partir de notre modèle d'interrogation flexible, sont plus rapides et plus précis par rapport à la recherche sur une base de données numérique.

Dans la dernière partie de notre modèle, notre contribution est de créer un modèle de classification supervisé médical, **Classifieur- RLR-CardF**, à partir des résumés linguistiques **RLR-CardF**. Notre but est d'établir un modèle de déduction, à l'aide d'informations

quantitatives et d'informations qualificatives, produites par nos résumés linguistiques. Nous avons utilisé pour cela le calcul de la similarité des variables linguistiques de résumé **RLR-CardF**.

Les méthodes de classifications supervisées existantes dans la littérature (SVM, kNN, RN,..) gèrent difficilement une grande masse de données(Big Data)où la complexité de calcul est élevée. Même les systèmes d'inférence flou trouve une difficulté lors de traitement d'une grande masse de données , cela est du au problème de l'augmentation les règles floues du système SIF.

Nous remarquons clairement que les résultats de notre modèle sont prometteurs et proche au langage naturel; il tient compte de trois critères importants: la réduction sémantique des données, le système d'interrogation flexible et le système de classification supervisée. Nous rappelons toujours le problème de volume de données rencontré dans les différentes méthodes de classification citées dans la littérature.

Notons aussi que le taux de classification obtenu dans notre travail est similaire à celui rencontré dans plusieurs travaux de l'état de l'art. Donc, les résultats obtenus sont dans l'ensemble encourageants.

Ce travail de thèse, nous permet d'ouvrir plusieurs perspectives de recherche :

- Amélioration des algorithmes proposés pour réduire la complexité de calcul et minimiser leur coût de calcul.
- Appliquer ce modèle sur d'autres bases de données médicales, avec des données hétérogènes ou sur un dossier médical complet.
- Intégrer ce modèle dans un système de classification en ligne.
- Combiner ce modèle avec d'autres méthodes de classification traditionnel(approche ensembliste).
- Traiter les problèmes des petits disjoints, des données manquantes, des données bruitées, des données chevauchées et des données frontières pour les ensembles de données binaires et multi-classes.
- Tester aussi le modèle proposé sur les données déséquilibrées.

Bibliographie

[1] L. A. Zadeh, “Fuzzy sets,” *Inf. Control*, vol. 8, no. 3, pp. 338–353, 1965.

[2] G. Raschia and N. Mouaddib, “SAINTETIQ: A fuzzy set-based approach to database summarization,” *Fuzzy Sets Syst.*, vol. 129, pp. 137–162, 2002.

[3] R. R. Yager, “A new approach to the summarization of data,” *Inf. Sci. (Ny).*, vol. 28, no. 1, pp. 69–86, 1982.

[4] P. Bosc, D. Dubois, O. Pivert, H. Prade, and M. De Calmes, “Fuzzy summarization of data using fuzzy cardinalities,” in *Proceedings of IPMU*, 2002, vol. 2002, pp. 1553–1559.

[5] L. Lietard, “A new definition for linguistic summaries of data,” in *Fuzzy Systems, 2008. FUZZ-IEEE 2008.(IEEE World Congress on Computational Intelligence). IEEE International Conference on*, 2008, pp. 506–511.

[6] D. A. Huffman, “A method for the construction of minimum-redundancy codes,” *Proc. IRE*, vol. 40, no. 9, pp. 1098–1101, 1952.

[7] W. A. Voglozin, “Le résumé linguistique de données structurées comme support pour l’interrogation,” Université de Nantes, 2007.

[8] L. Naoum, “Un modèle multidimensionnel pour un processus d’analyse en ligne de résumés flous,” Université de Nantes, 2006.

[9] Y. Peneveyre and C. Lambercy, “Compression de données sans pertes,” pp. 23–64.

[10] J. Han, Y. Fu, W. Wang, J. Chiang, W. Gong, K. Koperski, D. Li, Y. Lu, A. Rajan, N. Stefanovic, and others, “DBMiner: A System for Mining Knowledge in Large Relational Databases.,” in *KDD*, 1996, vol. 96, pp. 250–255.

[11] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh, “Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals,” *Data Min. Knowl. Discov.*, vol. 1, no. 1, pp. 29–53, 1997.

[12] H. V Jagadish, R. T. Ng, B. C. Ooi, and A. K. H. Tung, “Itcompress: An iterative semantic compression algorithm,” in *Data Engineering, 2004. Proceedings. 20th International Conference on*, 2004, pp. 646–657.

[13] S. P. Ghosh, “Statistical relational model,” in *International Conference on Scientific and Statistical Database Management*, 1988, pp. 338–355.

[14] O. Lebeltel, P. Bessière, J. Diard, and E. Mazer, “Bayesian robot programming,” *Auton. Robots*, vol. 16, no. 1, pp. 49–79, 2004.

[15] L. Fan, P. Cao, J. Almeida, and A. Z. Broder, “Summary cache: a scalable wide-area web cache sharing protocol,” *IEEE/ACM Trans. Netw.*, vol. 8, no. 3, pp. 281–293, 2000.

[16] M. Fran, M. Recherche, and E. L. Universit, “Résumés linguistiques de bases de données relationnelles,” *Sci. York*, pp. 1–14, 2007.

[17] R. R. Yager, “A new approach to the summarization of data,” *Inf. Sci. (Ny).*, vol. 28, no. 1, pp. 69–86, 1982.

[18] D. Dubois and H. Prade, “On data summarization with fuzzy sets,” *fifth IFSA World Congr.*, vol. 1, pp. 465–468, 1993.

[19] G. Raschia and N. Mouaddib, “SAINTETIQ: a fuzzy set-based approach to database summarization,” *Fuzzy sets Syst.*, vol. 129, no. 2, pp. 137–162, 2002.

[20] W. a. Voglozin, G. Raschia, L. Ughetto, and N. Mouaddib, “Querying a summary of database,” *J. Intell. Inf. Syst.*, vol. 26, no. 1, pp. 59–73, 2006.

[21] J. Kacprzyk, “Fuzzy logic for linguistic summarization of databases,” *IEEE Int. Fuzzy Syst. Conf. Proc.*, pp. 813–818, 1999.

[22] J. Kacprzyk, R. R. Yager, and S. Zadrożny, “A fuzzy logic based approach to linguistic summaries of databases,” *Int. J. Appl. Math. Comput. Sci.*, vol. 10, no. 4, pp. 813–834, 2000.

[23] D. Rasmussen and R. R. Yager, “A fuzzy SQL summary language for data discovery,” *Fuzzy Inf. Eng. A Guid. tour Appl.*, pp. 253–264, 1997.

[24] Z. Pei, Y. Xu, D. Ruan, and K. Qin, “Extracting complex linguistic data summaries from personnel database via simple linguistic aggregations,” *Inf. Sci. (Ny).*, vol. 179, no. 14, pp. 2325–2332, 2009.

[25] L. A. Zadeh, “A prototype-centered approach to adding deduction capability to search engines-the concept of protoform,” in *Fuzzy Information Processing Society, 2002. Proceedings. NAFIPS. 2002 Annual Meeting of the North American*, 2002, pp. 523–525.

[26] L. A. Zadeh, “A computational approach to fuzzy quantifiers in natural languages,” *Comput. Math. with Appl.*, vol. 9, no. 1, pp. 149–184, 1983.

[27] R. R. Yager, “Quantifier guided aggregation using OWA operators,” *Int. J. Intell. Syst.*, vol. 11, no. 1, pp. 49–73, 1996.

[28] L. Liétard, “A functional interpretation of linguistic summaries of data,” *Inf. Sci. (Ny).*, vol. 188, pp. 1–16, 2012.

[29] D. Rocacher and P. Bosc, “The set of fuzzy rational numbers and flexible querying,” *Fuzzy Sets Syst.*, vol. 155, no. 3, pp. 317–339, 2005.

[30] O. Pivert, A. Hadjali, and G. Smits, “Estimating the relevance of a data source using a fuzzy-cardinality-based summary,” in *Intelligent Systems (IS), 2010 5th IEEE International Conference*, 2010, pp. 96–101.

[31] R. Grin, “Introduction aux bases de données, modèle relationnel,” *Univ. Nice Sophia--antip.*, pp. 40–41, 2002.

[32] D. salah Eddine, “Intérrogation des bases de données dans le cluster sans sgbd,” 2013.

[33] Y. Loiseau, “Recherche flexible d ’ information par filtrage flou qualitatif,” vol. 33, no.

0, 2004.

[34] P. Bosc, M. Galibourg, and G. Hamon, “Fuzzy querying with SQL: extensions and implementation aspects,” *Fuzzy sets Syst.*, vol. 28, no. 3, pp. 333–349, 1988.

[35] P. Bosc and O. Pivert, “Some approaches for relational databases flexible querying,” *J. Intell. Inf. Syst.*, vol. 1, no. 3–4, pp. 323–354, 1992.

[36] P. Bosc and O. Pivert, “SQLf: a relational database language for fuzzy querying,” *Fuzzy Syst. IEEE Trans.*, vol. 3, no. 1, pp. 1–17, 1995.

[37] F. E. Petry, “Database Fundamentals,” in *Fuzzy Databases*, Springer, 1996, pp. 1–30.

[38] B. Larsen and C. Aone, “Fast and effective text mining using linear-time document clustering,” in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, 1999, pp. 16–22.

[39] H. Bouaziz, P. Narchi, F. J. Mercier, T. Labaille, N. Zerrouk, J. Girod, and D. Benhamou, “Comparison between conventional axillary block and a new approach at the midhumeral level,” *Anesth. Analg.*, vol. 84, no. 5, pp. 1058–1062, 1997.

[40] A. Motro, “BAROQUE: A browser for relational databases,” *ACM Trans. Inf. Syst.*, vol. 4, no. 2, pp. 164–181, 1986.

[41] A. Motro, “FLEX: A tolerant and cooperative user interface to databases,” *IEEE Trans. Knowl. Data Eng.*, vol. 2, no. 2, pp. 231–246, 1990.

[42] J. Fink and A. Kobsa, “A review and analysis of commercial user modeling servers for personalization on the world wide web,” *User Model. User-adapt. Interact.*, vol. 10, no. 2–3, pp. 209–249, 2000.

[43] T. Ichikawa and M. Hirakawa, “ARES: a relational database with the capability of performing flexible interpretation of queries,” *IEEE Trans. Softw. Eng.*, no. 5, pp. 624–634, 1986.

[44] A. Motro, “VAGUE: A user interface to relational databases that permits vague queries,” *ACM Trans. Inf. Syst.*, vol. 6, no. 3, pp. 187–214, 1988.

[45] M. Lacroix and P. Lavency, “Preferences; Putting More Knowledge into Queries.,” in *VLDB*, 1987, vol. 87, pp. 1–4.

[46] F. Herrera and E. Herrera-Viedma, “Aggregation operators for linguistic weighted information,” *IEEE Trans. Syst. Man, Cybern. Part ASystems Humans.*, vol. 27, no. 5, pp. 646–656, 1997.

[47] K. Abbaci, F. Lemos, A. Hadjali, D. Grigori, L. Lietard, D. Rocacher, and M. Bouzeghoub, “An Approach Based on Fuzzy Sets to Selecting and Ranking Business Processes,” *2011 IEEE 13th Conf. Commer. Enterp. Comput.*, pp. 213–218, 2011.

[48] P. Bosc, A. Hadjali, and O. Pivert, “An approach to competitive conditional fuzzy preferences in database flexible querying,” *2008 IEEE Conf. Intell. Syst.*, 2008.

[49] B. Bouchon-Meunier and Y. Jia, “Linguistic modifiers and imprecise categories,” *Int. J. Intell. Syst.*, vol. 7, no. 1, pp. 25–36, 1992.

[50] D. Dubois and H. Prade, “Fuzzy sets and statistical data,” *Eur. J. Oper. Res.*, vol. 25, no. 3, pp. 345–356, 1986.

[51] L. Ughetto, W. A. Voglozin, and N. Mouaddib, “Personalized database querying using data summaries,” in *Fuzzy Systems, 2006 IEEE International Conference on*, 2006, pp. 736–743.

[52] W. A. Voglozin, G. Raschia, L. Ughetto, and N. Mouaddib, “Interrogation de résumés de données et réparation de requêtes Querying data summaries and repairing queries Interrogation des résumés.”

[53] W. A. Voglozin, G. Raschia, L. Ughetto, and N. Mouaddib, “Querying the saintetiq summaries-dealing with null answers,” in *Fuzzy Systems, 2005. FUZZ’05. The 14th IEEE International Conference on*, 2005, pp. 585–590.

[54] L. Ughetto, W. a. Voglozin, and N. Mouaddib, “Database querying with personalized vocabulary using data summaries,” *Fuzzy Sets Syst.*, vol. 159, no. 15, pp. 2030–2046, 2008.

[55] J. Kacprzyk and S. Zadrożny, “Computing with words in intelligent database querying: standalone and Internet-based applications,” *Inf. Sci. (Ny).*, vol. 134, no. 1–4, pp. 71–109, 2001.

[56] W. Amenel, “R ´ esum ´ es de donn ´ ees pour la personnalisation de requˆ etes Personalized database querying using data summaries R ´ esum ´.”

[57] P. Badins, “Utilisation des cardinalités dans les résumés linguistiques de données Résumé,” pp. 1–29.

[58] C. J. Date, *Introducción a los sistemas de bases de datos*. Pearson Educación, 2001.

[59] L. A. Zadeh, “The concept of a linguistic variable and its application to approximate reasoning-I,” *Inf. Sci. (Ny).*, vol. 8, no. 4, pp. 301–357, 1975.

[60] J. Kacprzyk and S. Zadrozny, “Protoforms of Linguistic Database Summaries as a Tool for Human-Consistent Data Mining,” *14th IEEE Int. Conf. Fuzzy Syst. 2005. FUZZ ’05.*, pp. 591–596, 2005.

[61] R. R. Yager and F. E. Petry, “A multicriteria approach to data summarization using concept ontologies,” *IEEE Trans. Fuzzy Syst.*, vol. 14, no. 6, pp. 767–779, 2006.

[62] R. R. Yager, “Measures of specificity over continuous spaces under similarity relations,” *Fuzzy Sets Syst.*, vol. 159, no. 17, pp. 2193–2210, 2008.

[63] J. Kacprzyk and S. Zadrokny, “Linguistic Summarization of Data Sets Using Association Rules,” no. x, pp. 702–707.

[64] J. Kacprzyk, A. Wilbik, and S. Zadrozny, “Linguistic Summaries of Time Series via an OWA Operator Based Aggregation of Partial Trends,” *2007 IEEE Int. Fuzzy Syst. Conf.*,

no. xi, pp. 1–6, 2007.

[65] H. Iorxv and H. W. Txdqwlwpv, “Abilitation a,” 2005.

[66] D. Rocacher, I. I. U. T. Lannion, and L. Cédex, “Préférences et quantités dans le cadre de l ’ interrogation flexible : sur la prise en compte d ’ expressions quantifiées.”

[67] O. Pivert, “Contribution à l’interrogation flexible de bases de données: expression et évaluation de requêtes floues,” 1991.

[68] P. Bosc and L. Liétard, “A general technique to measure gradual properties of fuzzy sets,” in *International Conference of the Fuzzy Sets Association (IFSA ?5)*, 2005, pp. 485–490.

[69] D. Rocacher, “Relations d ’ ordre floues sur des quantités floues et expression de requêtes flexibles Fuzzy order relations on fuzzy quantities and flexible queries.”

[70] D. Dubois and H. Prade, “The mean value of a fuzzy number,” *Fuzzy sets Syst.*, vol. 24, no. 3, pp. 279–300, 1987.

[71] F. Díaz-Hermida and a. Bugarín, “Linguistic summarization of data with probabilistic fuzzy quantifiers,” *Proc. of ESTYLF’10*, pp. 255–260, 2010.

[72] L. Liétard, “A new definition for linguistic summaries of data,” *Proc. IEEE Int. Conf. Fuzzy Syst.*, pp. 506–511, 2008.

[73] L. Liétard, “A functional interpretation of linguistic summaries of data,” *Inf. Sci. (Ny).*, vol. 188, pp. 1–16, 2012.

[74] D. A. Zighed, Y. Kodratoff, and A. Napoli, “Extraction de connaissance {à} partir d?ne base de donn{é}es,” *Bull. AFIA*, vol. 1, 2001.

[75] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” *AI Mag.*, vol. 17, no. 3, p. 37, 1996.

[76] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons, 2011.

[77] C. Lurette, “D{é}veloppement d’une technique neuronale auto-adaptative pour la classification dynamique de donn{é}es {é}volutives: application {à} la supervision d’une presse hydraulique,” Lille 1, 2003.

[78] J. C. Bezdek, L. O. Hall, and L. Clarke, “Review of MR image segmentation techniques using pattern recognition,” *Med. Phys.*, vol. 20, no. 4, pp. 1033–1048, 1993.

[79] V. N. Vapnik and V. Vapnik, *Statistical learning theory*, vol. 1. Wiley New York, 1998.

[80] H. Larochelle, “Étude de techniques d’apprentissage non-supervisé pour l’amélioration de l’entra{\^\i}nement supervisé de modèles connexionnistes,” 2009.

[81] B. Fethi, B. Abdelhafid, B. Mortada, C. M. Amine, and S. Samira, “R econnaissance par N euro -F loue du D iab e ` te,” 2009.

[82] T. M. Mitchell and others, “Machine learning. 1997,” *Burr Ridge, McGraw Hill*, vol. 45, no. 37, pp. 870–877, 1997.

[83] A. Supervisé, “IGN,” Université Pierre et Marie Curie, Paris, 2001.

[84] M. S. Ahmed, “Analyse des méthodes d’apprentissage à base de noyaux: Application au diagnostic et à la classification des cellules cancéreuses,” Université Mohamed Boudiaf des sciences et de la technologie d’Oran, 2017.

[85] L. Henriet, “Systèmes d’évaluation et de classification multicritères pour l’aide à la décision: Construction de modèles et procédures d’affectation,” Université Paris Dauphine-Paris IX, 2000.

[86] C. Taylor, D. Michie, and D. Spiegalhalter, “Machine Learning, Neural and Statistical Classifiers,” 1994.

[87] L. Candillier, “Contextualisation, visualisation et évaluation en apprentissage non supervisé,” Université Charles de Gaulle-Lille III, 2006.

[88] C. Salperwyck, “Apprentissage incr ´ emental en-ligne sur flux de donn ´ ees par,” 2012.

[89] S. M. Weiss and C. A. Kulikowski, *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. Morgan Kaufmann Publishers Inc., 1991.

[90] V. N. VAPNIK., “The nature of statistical learning theory,” *Springer-Verlag*, 1995.

[91] J. C. Platt, “12 fast training of support vector machines using sequential minimal optimization,” *Adv. kernel methods*, pp. 185–208, 1999.

[92] T. Joachims, “Estimating the generalization performance of a SVM efficiently,” 2000.

[93] N. Cristianini and J. Shawe-Taylor, “An introduction to support vector machines.” Cambridge University Press Cambridge, 2000.

[94] R. Herbrich, *Learning kernel classifiers: theory and algorithms*. Mit Press, 2001.

[95] S. Abe, *Support vector machines for pattern classification*, vol. 2. Springer, 2005.

[96] S. V. Machine and C. Roc, “Classification supervisée Aperçu de quelques méthodes avec le logiciel R,” pp. 1–20.

[97] Z. Hakima, “Thème Clustering par fusion floue de données appliqué à la segmentation d ’ images IRM,” 2008.

[98] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, “Classification and regression trees. Wadsworth & Brooks,” *Monterey, CA*, 1984.

[99] J. R. Quinlan, “Induction of decision trees,” *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[100] J. R. Quinlan, “C4. 5: Programs for Empirical Learning Morgan Kaufmann,” *San Fr. CA*,

1993.

[101] G. V Kass, “An exploratory technique for investigating large quantities of categorical data,” *Appl. Stat.*, pp. 119–127, 1980.

[102] T. Mitchell, “Machine Learning. MacGraw-Hill Companies,” *Inc*, vol. 1, p. 997, 1997.

[103] C. Scott and R. Nowak, “Dyadic classification trees via structural risk minimization,” in *NIPS*, 2002, pp. 359–366.

[104] G. Lebrun, “S ´ election de mod ` eles pour la classification supervis ´ ee a Vaste Marge). Application en traitement et analyse d ’ images . To cite this version : Présentée par Gilles Lebrun DOCTORAT de l ’ UNIVERSITE de CAEN Spécialité : Informatique Sélection de,” 2006.

[105] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in the brain.,” *Psychol. Rev.*, vol. 65, no. 6, p. 386, 1958.

[106] S. Haykin, “Neural Networks: A Comprehensive Foundation.. Tom Robbins--Prentice Hall.” Inc, 1999.

[107] G. Deyfus and J. M. Martinez, “Les R{é}seaux de Neurones Methodologies et Applications,” *Eyrolle,* 2004.

[108] S. Tufféry, *Data mining et statistique d{é}cisionnelle: l’intelligence dans les bases de donn{é}es*. Editions Technip, 2005.

[109] S. Tufféry, *Data mining et scoring: bases de donn{é}es et gestion de la relation client*. Dunod, 2002.

[110] J. C. Bezdek, “Pattern Recognition with Fuzzy Objective Algorithms Plenum Press New York Google Scholar,” 1981.

[111] L. Khodja, “Contribution {à} la classification floue non supervis{é}e,” 1997.

[112] M. S. Kamel and S. Z. Selim, “A thresholded fuzzy c-means algorithm for semi-fuzzy clustering,” *Pattern Recognit.*, vol. 24, no. 9, pp. 825–833, 1991.

[113] L. A. Zadeh, “Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic,” *Fuzzy sets Syst.*, vol. 90, no. 2, pp. 111–127, 1997.

[114] A. Wilbik and J. M. Keller, “A distance metric for a space of linguistic summaries,” *Fuzzy Sets Syst.*, vol. 208, pp. 79–94, 2012.

[115] J. Kacprzyk and A. Wilbik, “Towards an efficient generation of linguistic summaries of time series using a degree of focus,” in *Fuzzy Information Processing Society, 2009. NAFIPS 2009. Annual Meeting of the North American*, 2009, pp. 1–6.

[116] D. Pilarski, “Linguistic summarization of databases with quantirius: a reduction algorithm for generated summaries,” *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.*, vol. 18, no. 03, pp. 305–331, 2010.

[117] R. J. Almeida, M. Lesot, B. Bouchon-meunier, G. Moyse, R. J. Almeida, M. Lesot, B.

Bouchon-meunier, U. Kaymak, G. Moyse, R. J. Almeida, M. Lesot, B. Bouchon-meunier, U. Kaymak, and G. Moyse, “Septic Shock Patient Data To cite this version : HAL Id : hal-00932850 Linguistic Summaries of Categorical Time Series for Septic Shock Patient Data,” 2015.

[118] P. Jaccard, “The distribution of the flora in the alpine zone.,” *New Phytol.*, vol. 11, no. 2, pp. 37–50, 1912.

[119] B. De Baets and R. Mesiar, “Pseudo-metrics and T-equivalences,” *J. Fuzzy Math. 5 (1997), 471-481*, 1997.

[120] B. De Baets and R. Mesiar, “Metrics and T-equalities,” *J. Math. Anal. Appl.*, vol. 267, no. 2, pp. 531–547, 2002.

[121] A. Wilbik, J. M. Keller, and G. L. Alexander, “Similarity evaluation of sets of linguistic summaries,” *Int. J. Intell. Syst.*, vol. 27, no. 10, pp. 926–938, 2012.

[122] J. Kacprzyk, M. Fedrizzi, and H. Nurmi, “Group decision making and consensus under fuzzy preferences and fuzzy majority,” *Fuzzy Sets Syst.*, vol. 49, no. 1, pp. 21–31, 1992.

[123] M. Fedrizzi, J. Kacprzyk, and others, “Soft consensus measures for monitoring real consensus reaching processes under fuzzy preferences,” *Control Cybern.*, vol. 15, no. 3–4, pp. 309–323, 1986.

[124] R. R. Yager, “Ordered weighted averaging aggregation operators in multi-criteria decision making, IEEE Trans. on Systems, Man and Cybernetics, 18,” 1988.

[125] T. Murofushi and M. Sugeno, “Fuzzy measures and fuzzy integrals,” *Fuzzy Meas. Integr. Theory Appl.*, pp. 3–41, 2000.

[126] G. J. Klir and Z. Wang, “Fuzzy measure theory.” Plenum Press, New York, 1992.

[127] J. Keller, P. Gader, and A. K. Hocaoglu, “Fuzzy integrals in image processing and recognition,” *Fuzzy Meas. Integr. Theory Appl.*, pp. 435–466, 2000.

[128] M. J. Rantz, R. T. Porter, D. Cheshier, D. Otto, C. H. Servey III, R. A. Johnson, M. Aud, M. Skubic, H. Tyrer, Z. He, and others, “TigerPlace, a State-Academic-Private project to revolutionize traditional Long-Term care,” *J. Hous. Elderly*, vol. 22, no. 1–2, pp. 66–85, 2008.

[129] E. Hanisch, R. Brause, B. Arlt, J. Paetz, and K. Holzer, “The MEDAN database,” *Comput. Methods Programs Biomed.*, vol. 75, no. 1, pp. 23–30, 2003.

[130] K. Kaczmarek and O. Hryniewicz, “Time Series Classification with Linguistic Summaries,” *Proc. of EUSFLAT’15*, pp. 1–7, 2015.

[131] A. Asuncion and D. Newman, “UCI machine learning repository.” 2007.

[132] M. Delgado, D. Sánchez, and M. A. Vila, “Fuzzy cardinality based evaluation of quantified sentences,” *Int. J. Approx. Reason.*, vol. 23, no. 1, pp. 23–66, 2000.

[133] D. Amghar and M. A. Chikh, “The summary linguistic and medical database,” in

Proceedings of the International Conference on Intelligent Information Processing, Security and Advanced Communication, 2015, p. 33.

[134] M. Sekkal and M. A. Chikh, “NEURO◆GENETIC APPROACH TO CLASSIFICATION OF CARDIAC ARRYTHMIAS,” *J. Mech. Med. Biol.*, vol. 12, no. 01, p. 1250010, 2012.

[135] M. El, A. Lazouni, M. El, H. Daho, and N. Settouti, “SVM Computer Aided Diagnosis for Anesthetic Doctors,” no. 5, pp. 235–240, 2013.

[136] D. Amghar and M. A. Chikh, “Extracting a Linguistic Summary from a Medical Database,” *Int. J. Intell. Syst. Appl.*, no. December, pp. 16–26, 2018.

[137] R. J. Almeida, M.-J. Lesot, B. Bouchon-Meunier, U. Kaymak, and G. Moyse, “Linguistic summaries of categorical time series for septic shock patient data,” in *Fuzzy Systems (FUZZ), 2013 IEEE International Conference on*, 2013, pp. 1–8.

[138] L. Gonzalez-Abril, H. Nuñez, C. Angulo, and F. Velasco, “GSVM: An SVM for handling imbalanced accuracy between classes inbi-classification problems,” *Appl. Soft Comput.*, vol. 17, pp. 23–31, 2014.

[139] Y.-H. Shao, W.-J. Chen, J.-J. Zhang, Z. Wang, and N.-Y. Deng, “An efficient weighted Lagrangian twin support vector machine for imbalanced data classification,” *Pattern Recognit.*, vol. 47, no. 9, pp. 3158–3167, 2014.

[140] J. Guzaitis, A. Verikas, A. Gelzinis, and M. Bacauskiene, “A framework for designing a fuzzy rule-based classifier,” in *International Conference on Algorithmic DecisionTheory*, 2009, pp. 434–445.

[141] N. Settouti, M. Saidi, and M. A. Chikh, “Interpretable Classifier of Diabetes Disease,” vol. 4, no. 3, 2012.

[142] M. Blachnik and W. Duch, “Prototype-based threshold rules,” in *International Conference on Neural Information Processing*, 2006, pp. 1028–1037.

[143] K. J. Wang and A. M. Adrian, “Breast cancer classification using hybrid synthetic minority over-sampling technique and artificial immune recognition system algorithm,” *Int J Comput Sci Electron Eng*, vol. 1, no. 3, pp. 408–412, 2013.

[144] M. R. Hassan, M. M. Hossain, R. K. Begg, K. Ramamohanarao, and Y. Morsi, “Breast-cancer identification using HMM-fuzzy approach,” *Comput. Biol. Med.*, vol. 40, no. 3, pp. 240–251, 2010.

[145] I. Maglogiannis, E. Zafiropoulos, and I. Anagnostopoulos, “An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers,” *Appl. Intell.*, vol. 30, no. 1, pp. 24–36, 2009.

[146] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, 2005.

[147] Q. Cheng, H. Zhou, and J. Cheng, “The fisher-markov selector: fast selecting maximally separable feature subset for multiclass classification with applications to high-

dimensional data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1217–1233, 2011.

[148] D. Lin and X. Tang, "Conditional infomax learning: an integrated framework for feature extraction and fusion," in *European Conference on Computer Vision*, 2006, pp. 68–82.

[149] R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern Classification, New York: Wiley Interscience," 2000.

[150] H. H. Yang and J. Moody, "Data visualization and feature selection: New algorithms for nongaussian data," in *Advances in Neural Information Processing Systems*, 2000, pp. 687–693.

Résumé

D'une manière générale, la collecte des données médicales créent une grande base des données numériques. Notre travail de recherche vise à extraire des résumés linguistiques, à base de calcul de cardinalité floue. Ce type de résumé permet de construire une base de connaissances réduite qui contient toutes les informations essentielles, pour une meilleure décision. Cette dernier est aussi utilisé pour construire un système interrogation flexible qui traite des requêtes sémantiques simples et des requêtes complexes en langage naturel ,à la l'aide d'une nouvelle approche, qui est proposé dans notre thèse de recherche. Aussi de développer un nouveau type de classifieur supervisé basé sur ces résumés linguistiques des données médicales. Pour réaliser cet objectif, nous utilisons le calcul de la similarité entre les différents ensembles flous de résumés linguistiques. Les solutions proposées ont été validées expérimentalement sur des bases de données médicales réelles. Les résultats obtenus sont comparés à ceux de l'état de l'art où nous montrons l'efficacité de modèle de la déduction proposée.

Mots-clés: Données numériques médicales, résumé linguistique, requêtes floues, classification des données médicales, logique floue.

Abstract

In general, the collection of medical data creates a large database of digital data. Our research aims to extract language summaries, based calculates fuzzy cardinality. This kind of summary helps build a small knowledge base that contains all the essential information for a better decision. The latter is also used to build a flexible query system that deals with simple semantic queries and complex queries in natural language using a new approach, proposed in our research thesis. Also to develop a new type of supervised classifier based on these linguistic summaries of medical data. To achieve this objective, we use the calculation of the similarity between the different fuzzy sets of linguistic summaries. The proposed solutions have been validated experimentally on real biomedical databases. The results obtained are compared with those of the state of the art where we show the model efficiency of the proposed deduction.

Keywords: Medical digital data, linguistic summary, fuzzy queries, medical data classification, fuzzy logic.

الملخص

بشكل عام ، يخلق جمع البيانات الطبية قاعدة بيانات كبيرة من البيانات الرقمية. يهدف عملنا البحثي إلى استخراج الملخصات اللغوية على أساس حساب الأصالة الغامض. يساعد هذا النوع من الملخص في بناء قاعدة معارف صغيرة تحتوي على جميع المعلومات الأساسية لاتخاذ قرار أفضل. يستخدم هذا الأخير أيضًا لبناء نظام الاسترجاع مرن يتعامل مع الطلبيات الدلالية والطلبيات المعقدة باللغة الطبيعية باستخدام منهج جديد ، مقترح في أطروحة البحث. أيضا لتطوير نوع جديد من المصنف الخاضع للإشراف على أساس هذه الملخصات اللغوية للبيانات الطبية. لتحقيق هذا الهدف، نستخدم حساب التشابه بين مجموعات مختلفة من الملخصات اللغوية. تم التحقق من صحة الحلول المقترحة تجريبيًا على قواعد البيانات الطبية الحيوية الحقيقية. تتم مقارنة النتائج التي تم الحصول عليها مع تلك التي تم الحصول عليها من أحدث ما توصلنا إليه من فعالية نموذجية للخصم المقترح.

الكلمات المفتاحية: بيانات رقمية طبية ، ملخص لغوي ، استفسارات غامضة ، تصنيف البيانات الطبية ، المنطق الضبابي.