Ministère de l'Enseignement Supérieur
et de la Recherche Scientifique
Université de Tlemcen
Faculté de Science
Département Informatique

MÉMOIRE DE FIN D'ÉTUDE

pour obtenir le grade de

Master en Informatique

Spécialité : **MID**

présenté et soutenu publiquement

par

Md. Kalache Soumia

Melle. Kouloughli Asma

le 02 Juillet 2013

Titre:

Les Forêts Aléatoires Floues

Jury

Président du jury.	Mr. Hadjila Fethallah,	MAA UABB Tlemcen
Examinateur.	Mr. Benazzouz Mourtada,	MAA UABB Tlemcen
Examinateur.	Mr. Belabed Amine,	MAA UABB Tlemcen
Directeur de mémoire.	Pr. Chikh Mohamed Amine,	Professeur UABB Tlemcen
Co-Directeur de mémoire.	Melle Settouti Nesma,	MAB à UABB Tlemcen

Nous dédions ce travail à :

Nos parents,

Nos maris,

Nos grands parents,

Toute la famille,

Nos amis,

Qu'ils trouvent ici l'expression de toute notre reconnaissance.

Remerciements

Nous tenons tout d'abord à remercier Dieu le tout puissant et miséricordieux, qui nous a donné la force et la patience d'accomplir ce modeste travail.

En second lieu, nous tenons à remercier notre encadreur Mr CHIKH M.A Professeur à l'Univérsité de Tlemcen, de ses précieux conseils et son aide durant toute la période du travail.

Nos vifs remerciment à Melle SETTOUTI N. Maitre Assistant à l'Univérsité de Tlemcen, en tant que co-encadreur et lui témoigner notre gratitude pour sa patience et son soutien qui nous a été précieux afin de mener notre travail à bon port.

Nous tenons à remercier aussi Mr HADJILA F. pour l'intérêt porté à ce travail et d'avoir accepté de présider ce jury.

En suite nous désirons adresser nos sincères sentiments à Mr BENAZZOUZ M. et BELABED A. d'avoir éxaminé ce travail.

Enfin, nous tenons également à remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.

Résumé

Nous traitons dans ce mémoire l'extraction de la connaissance à partir des données, en utilisant les forêts aléatoires floues qui combinent la robustesse des arbres de décision, la puissance du caractère aléatoire qui augmente la diversité des arbres dans la forêt, et la flexibilité de la logique floue. Ils ont la spécificité de contrôler des données imparfaites, de réduire le taux d'erreurs et de mettre en évidence plus de robustesse et plus d'interprétabilité.

Dans le cadre de notre travail nous nous intéresserons a la construction une forêt d'arbres de décision floues (de types Fuzzy CART) pour la classification de données médicales, nous optimisons ensuite ces arbres avec l'algorithme Fuzzy C-Mean qui nous permettra une meilleure répartition des données et ainsi qu'une régularisation des contraintes qui s'appliquent sur les paramètres des fonctions d'appartenance floues. Cet algorithme réduit le nombre de sous-ensembles flous et minimise le nombre de règles pour une connaissance ciblée.

Mots clés

Forêt aléatoire, arbre de décision, logique floue, connaissances.

Abstract

We deal in this paper the extraction of knowledge from data, using fuzzy random bits that combine the robustness of decision trees ; the power of randomness increases the diversity of trees in the forest, and flexibility fuzzy logic. They control the specificity of imperfect data, reduce the error rate and highlight more robustness and interpretability.

In our work we focus on the construction of a fuzzy decision tree forest (types Fuzzy CART) for the classification of medical data, we then optimize these trees with Fuzzy C-Mean algorithm that will allow us to better distribution of data and a regularization of the constraints that apply to the parameters of fuzzy membership functions. This algorithm reduced the number of fuzzy sets and minimizes the number of rules for a specific knowledge.

Keywords

Random forest, decision tree, fuzzy logic, knowledge.

Table des matières

Table des figures

Liste des tableaux

Liste des abreviations

Bagging : Boostrap and agregating.

CART : Classification And Regression Tree.

FCART : Fuzzy Classification And Regression Tree.

FCM : Fuzzy C-Means.

FID : Division The Inspection of the Forest.

FRF : Fuzzy Random Forest.

ID3 : Induction of Decision Tree.

OOB : Out Of Bag.

RF : Random Forest.

TC : Taux de Classification.

TDIDT : Top Down Induction of Deciosn Tree.

TSK : Takagi Segenou Kang.

UCI : University Californie Irvine.

Introduction générale

Pour certains domaines d'application, il est essentiel de produire des procédures de classification compréhensibles par l'utilisateur. C'est en particulier le cas pour l'aide au diagnostic médical où le médecin doit pouvoir interpréter les raisons du diagnostic.

A ce titre plusieurs travaux sont effectues afin de développer des outils d'aide au diagnostic et de classification des maladies, les arbres de décision répondent à ces contraintes car ils représentent graphiquement un ensemble de règles et sont aisément interprétables mais leur instabilité nous a ramène a élabore Des méthodes comme le Bagging (pour Bootstrap Aggregating), les Random Forests de Breiman permettent dans une certaine mesure de remédier à ce problème. Malgré la puissantes et la performance de ce dernier mais a la notion de stabilité n'est pas trop présente.

Dans ce mémoire de fin d'étude, nous nous intéressons à l'étude de la performance de l'algorithme Fuzzy Random Forest a decision de type fuzzy cart Nous remarquons que ce dernier donne des résultats assez médiocres avec un temps d'exécution très grand malgré que la stabilité recherché soit bien présente.

Nous proposons l'algorithme d'optimisation le Fuzzy C-Means qui a déjà montrer son intérêt d'application dans d'autres travaux au préalable pour l'apprentissage structurelle avec des paramètres floue. Cet algorithme réduire le nombre de sous-ensembles flous et de minimiser le nombre de règles pour une connaissance ciblée de fuzzy inferance système.

Notre plan de mémoire se compose de :

- **Chapitre I :** présentation de l'algorithme de base des forêts aléatoires en trois parties :
  1. Partie1 : l'arbre de décision classique.
  2. Partie 2 : les méthodes d'ensemble.
  3. Partie 3 : les forêts aléatoires (Ramdom Forest).

- **Chapitre II :** introduction aux forets aléatoires flous en présentant un état de l'art sur les travaux réalisés dans ce contexte.

- **Chapitre III :** résultats et interprétations des algorithmes implémentés avec une synthèse sur les différentes techniques.

En dernier lieu, une conclusion générale et les perspectives à venir dans ce travail.

Chapitre 1

Présentation de Random Forest

Introduction

Ce chapitre comporte trois parties : dans un premier temps, nous présentons un arbre de décision classique, et ses algorithmes de développement comme CART, la deuxième partie présente les méthodes d'ensemble et la troisième partie concerne les forêts aléatoires (Ramdom Forest).

1 Arbre de décision

1.1 Principe

Un arbre de décision est une méthode de classification et de prédiction la plus populaire en apprentissage supervisé, elle est simple à utiliser et à interpréter tout en gardant des performances très acceptables, elle donne des résultats satisfaisants pour des problèmes complexes (forêts aléatoires).

Il est largement répandu dans les domaines de statistiques, de l'ingénierie, de la théorie de la décision, ou encore de l'apprentissage automatique [Ber09].

Il existe plusieurs algorithmes de construction d'arbres, nous citons en particulier trois algorithmes connus :

1.2 Les algorithmes de construction d'arbre de décision

CART Classification and regression Tree

Algorithme CART signifie Classification and Regression Tree, il désigne une méthode statique élaboré par L. Breiman, J.H. Friedman, R.A. Olshen et C.J. Stone en 1984 [Aur06], qui construit des prédicateurs par arbre aussi bien en régression qu'en classification, il utilise un arbre binaire (toujours deux branches par nœud non feuilles).

Le principe est de diviser récursivement et le plus efficacement possible les exemples de l'ensemble d'apprentissage par des tests définis a l'aide des attributs jusqu'a ce que l'on obtienne des sous-ensembles d'exemples ne contenant (presque) que des exemples appartenant a une même classe.

Algorithme ID3 (Induction of Decision Tree)

ID3 est un algorithme de classification supervisé développée par Quinlan en 1983, il permet de remplacer les experts dans la construction d'un arbre de décision.

Le choix du test associé à un noeud se fait à l'aide de la fonction Gain. La méthode peut prendre en compte le cas où les valeurs de certains attributs sont non spécifiées. Elle prend également en compte le problème des attributs continus. On peut choisir entre arbres et règles, l'élagage se fait sur l'arbre ou sur le système de règles et se base sur une estimation de l'erreur réelle à partir de l'ensemble d'apprentissage.

Algorithme C4.5

C4.5 est un algorithme modifié de ID3 (implémentation plus facile), publié par Ross Quinlan en 1993.Cette méthode permet de travailler à la fois avec des données discrètes et des données continue [Beh09].

Contrairement à ID3, C4.5 est parfaitement réalisable dans des applications industrielles.

2 Méthodes d'ensemble

Le principe général des méthodes d'ensemble est de construire une collection de prédicateurs, pour ensuite agréger l'ensemble de leurs prédictions [Gen10].

- Dans le cadre de régression : agréger les prédictions de q prédicateurs revient par exemple à en faire la moyenne, chaque prédicateur fournit un y1 et la prédiction finale est alors :

$$\frac{1}{q}\sum_{l=1}^{n} y_1 \tag{1.1}$$

- Dans le cadre de classification : l'agrégation revient par exemple à faire un vote majoritaire parmi les classes fournies par les prédicteurs.

Les méthodes d'ensemble génèrent plusieurs règles de prédiction et mettent ensuite en commun leurs différentes réponses. L'heuristique de ces méthodes est que le prédicateur final soit meilleur que chacun des prédicteurs individuels. L'heuristique expliquant le succès de ces méthodes d'ensemble se résume ainsi [Gen10] :

- Chaque prédicteur individuel doit être relativement bon.
- Les préditeurs individuels doivent être différents les uns des autres.

Les méthodes d'ensemble se reposent sur une construction aléatoire d'une famille de modèle ''Bagging" et sur une construction adaptative, déterministe ou aléatoire d'une famille de modèle '' Boosting".
Il y'a deux types de méthodes d'ensemble :

1. **Les méthodes d'ensemble hétérogène :**
   Combinent un ensemble d'hypothèses $h_1 \dots h_T$, des algorithmes différents $L_1 \dots L_T$ sur un même ensemble d'apprentissage A.
2. **les méthodes d'ensemble homogène :**
   Combinent un ensemble d'hypothèses $h_1 \dots h_T$, produites par un même algorithme $L_1 \dots L_T$ sur un différent ensemble d'apprentissage A. Elles utilisent des stratégies adaptatives (boosting) ou aléatoires (bagging).

   Dans notre travail nous nous intéressons juste aux méthodes d'ensemble homogène comme :

2.1 Bagging

Le mot «bagging» est la contraction des mots «Bootstrap et Aggregating» [Ber09], inventé par Breiman en 1996 il est efficace pour corriger le manque de robustesse des arbres de décision.

Il consiste à la construction d'une règle de base sur m échantillons boostrap différents on en modifie les prédictions, et donc on construit à terme une collection de prédicteurs variés, qui sont ensuite agrégés par un vote (majoritaire des résultats des modèles) ou une moyenne des estimations c'est-à-dire que cette étape permet d'obtenir un prédicteur performant et a combiner plusieurs calssifieurs.

Boostrapping

Un bootstrap d'un ensemble T est l'ensemble obtenu en tirant $|T|$ fois des éléments de T uniformément au hasard et avec remise. Le bootstrapping d'un ensemble d'entraînement T produit un nouvel ensemble T' qui présente en moyenne $1-e^{-1} \approx 63\%$ instances uniques différentes de T quand $|T| >> 1$ [Sté11] (la création de plusieurs bag a partir d une base de donnée) .

Agrégation

On produit plusieurs bootstraps $T'_1, \ldots, T'_m$, chaque bootstrap T'_i étant utilisé pour entraîner un prédicteur t_i (penser ici à un arbre de régression, mais la technique s'applique à n'importe quelle famille de prédicteurs). Étant donnée une instance (x, y), on fait régresser chaque arbre, ce qui nous donne un ensemble de valeurs $y_1, \ldots, y_m$ prédites. Celles-ci sont alors aggrégées en calculant leur moyenne [Sté11] :

$$y = \frac{1}{m} \sum i \ y_i. \tag{1.2}$$

La principale force du bagging est donc de réduire l'instabilité pour augmenter les performances en généralisation. Mais il y a un autre point qui fait la force du bagging, ce sont les mesures out-of-bag. Nous expliquons maintenant cet outil très pratique en classification.

Exemple

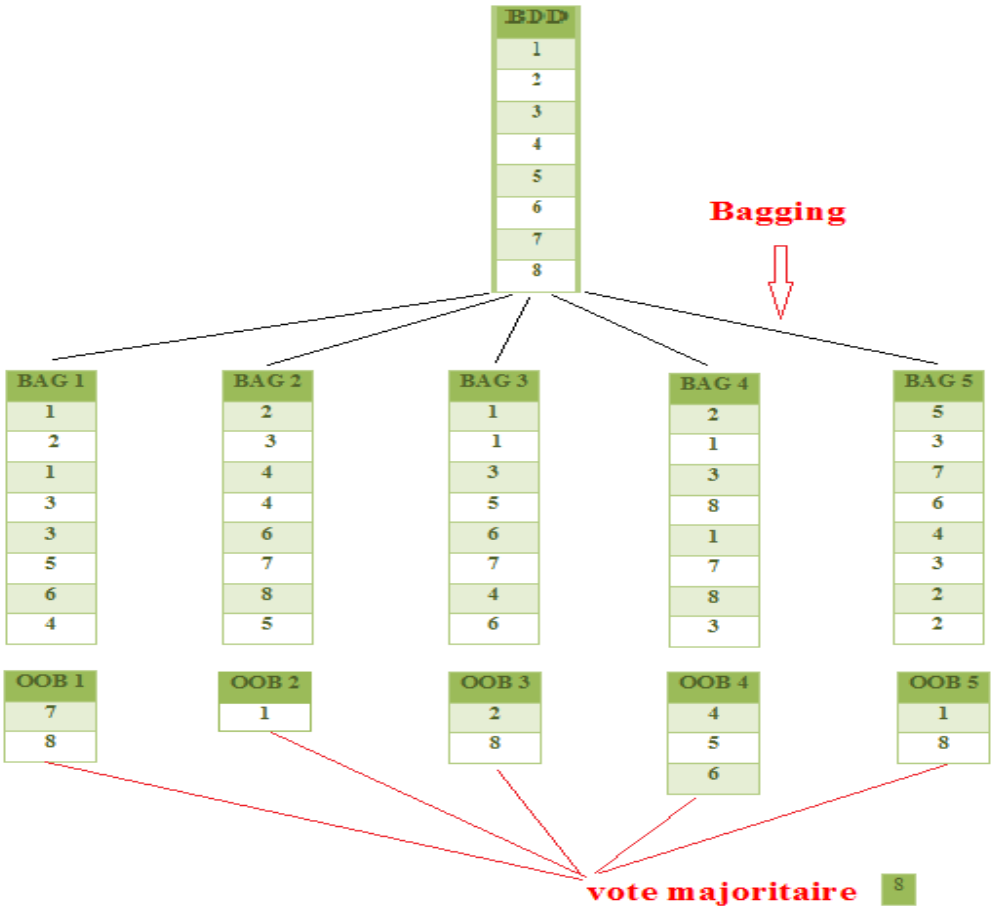

FIGURE 1.1 – Illustration du principe de Bagging pour un ensemble d'arbres de décision.

Mesures Out-Of-Bag (en dehors du bootstrap)

On prend une Base d'apprentissage A de m données, si l'on effectue un tirage aléatoire avec remise des données, il n'est pas impossible de tirer plusieurs fois la même donnée pour le même échantillon bootstrap. Sachant que pour chaque échantillon bootstrap 63,2% des exemples sont uniques de A, le reste étant des doublons .Ce résultat est intéressant car il indique que chaque classifieur n'apprend qu'une partie des données. Les autres données, celles qu'il ne connait pas, sont appelées les données out-of-bag (OOB), Dans le cadre d'un ensemble de classifieurs par exemple, si on souhaite estimer l'erreur en généralisation via un vote à majoritaire, il suffit de classer chaque donnée d'apprentissage de A en ne laissant voter que les classifieurs élémentaires pour les quels elle fait partie de l'ensemble out-of-bag. Selon Breiman

un grand nombre d'outils sont basés sur l'utilisation des forêts aléatoires, et beaucoup d'entre eux utilisent les données out-of-bag pour éviter d'avoir recours à un ensemble de données de validation [Ber09].

2.2 Boosting

Le Boosting est l'une des méthodes d'ensemble les plus performantes à ce jour, introduit par Freund et Shapire en 1996 [Gen10], c'est une version adaptative du Baggin dont le principe de départ est le même en donnant plus de poids lors de l'étape suivante, aux observations mal prédites.

Son but est d'améliorer la performance de l'algorithme d'apprentissage et créer progressivement les méthodes d'ensemble.

Les deux méthodes d'ensemble évoqués ont un principe général commun, il s'agit de partir d'une règle de prédiction de base puis perturber cette base de règle.
Nous détaillons à présent les forêts aléatoires (Random Forest) qui est une amélioration du Bagging spécifique aux modèles définis par des arbres binaire (CART).

3 Random Forest (Forêt aléatoire)

Random Forest (RF) traduit en français les forêts aléatoires introduits par Breiman en 2001, qui a définit les forêts aléatoires comme une famille des méthodes d'ensemble [BHAO09].

C'est une méthode statistique non-paramétrique et d'agrégation d'une collection d'arbre aléatoire (pour générer de la diversité dans les ensembles d'arbres se sont multipliées). Pour bien comprendre les forêts aléatoires il est intéressant de commencer par la définition publié dans l'article de Breiman.

3.1 Définition

Une Forêt Aléatoire est un classifieur constitué d'un ensemble de classifieurs élémentaires de type arbres de décision binaire dans le quel a été introduit de l'aléatoire [Sco12].

Soit $h(O1) \dots h(Oq)$ une collection de prédicteurs par arbre, où $(O_1 \dots O_q)$ est

une suite de variables aléatoires indépendante de l'échantillon d'apprentissage Ln. Le prédicteur des forêts aléatoires est obtenu par agrégation de cette collection de prédicteurs.

Le terme forêt aléatoire vient du fait que les prédicteurs individuels sont, ici, explicitement des prédicteurs par arbre, et du fait que chaque arbre dépend d'une variable aléatoire supplémentaire (c'est-à-dire en plus de L_n).

Une forêt aléatoire est l'agrégation d'une collection d'arbres aléatoires, elle peut par exemple être construite en générant des sous-ensembles aléatoires de caractéristiques pour chaque arbre et en générant des sous-ensembles aléatoires de données d'apprentissage pour chaque arbre (comme dans la méthode de Bagging).

RF génère un jeu d'arbre doublement perturbés ou chaque arbre du jeu est ainsi génère au départ d'un sous échantillon boostrap du jeu d'apprentissage complet.

De nombreux modèle de forêts aléatoires ont été crées qui correspondent a autant de manière d'incorporer de l'aléatoire dans les arbres on a pour exemple :

- **Le Tree Baggin :** [Sco12] introduit de l'aléatoire dans l'échantillon initial sélectionnant certains points plutôt que d'autres et laisse grandir l'arbre jusqu'à ce que chaque nœud comporte un unique élément.
- **Le Random Subspace :** [Sco12] consiste à sélectionner à chaque nœud K variables de manière aléatoire et parmi celles-ci à choisir celle qui minimise un certain critère.
- **La Random Forest :** [Sco12] qui consiste a mélangé le CART sans élagage, le bagging et le Random Subspace pour chaque arbre, on tire un échantillon a partir de l'échantillon initial, a chaque nœud, on choisit aléatoire K variables et on prend, parmi celle-ci, celle qui minimise le critère de l'algorithme CART. On laisse grandir l'arbre jusqu'à ce qu'il n'y ait plus qu'un seul élément dans chaque nœud.

3.2 Autres points importants

Il y a quelques autres points qui est intéressant de les cites.

1. **L'opérateur de combinaison :**

Cet opérateur utilise le vote à majoritaire chaque arbre participe au vote de la classe la plus populaire pour une donnée d'entrée x.

2. **La formalisation de la randomisation :**

   Des algorithmes d'induction de forêts aléatoires : elle se traduit par l'introduction d'un vecteur aléatoire dans la définition des classifieurs alors une famille de vecteurs aléatoires indépendants et identiquement distribués. Cela signifie dans un premier temps que durant le processus d'induction des forêts aléatoires, chaque arbre est induit indépendamment des autres arbres de décision déjà ajoutés à la forêt Ensuite, cela signifie que les principes de randomisation utilisés doivent être identiques pour l'induction de tous les arbres de la forêt. Selon Breiman [Bre01] la consistance des forêts sont fortement basés sur cette hypothèse d'indépendance et de distribution homogène. Il n'existe d'ailleurs à notre connaissance aucun algorithme d'induction de forêts aléatoires qui sort de ce cadre.

3. **Recherche exhaustive de K :**

   Le paramètre K fixe le nombre de caractéristiques sélectionnées aléatoirement à chaque noeud au cours de la procédure d'induction d'un arbre. Sa valeur est donc choisie dans l'intervalle $[1 \dots M]$, où M représente la dimension de l'espace de description et par défaut :

   K= $\sqrt{M}$ sachant que plus la valeur de K est petite et plus on introduit l'aléatoire.

   Pour résumer, la foret d'arbre de décision aleatoire est construit selon la procédure suivante [BHAO09] :

   - Pour N données de l'ensemble d'apprentissage, tirer aléatoirement N individus avec remise. L'ensemble résultant l'induction de l'arbre en question.
   - Pour M caractéristiques, un nombre K « M est spécifié de sorte qu'à chaque noeud de l'arbre, un sous ensemble de K caractéristiques soit tiré aléatoirement, parmi lesquelles la meilleure est ensuite sélectionnée pour le partitionnement.
   - L'arbre est ainsi construit jusqu'a l'obtention de sa taille maximale.
   - Aucun élagage n'est réalisé.

Conclusion

Les arbres de décision tels que nous venons de les aborder présentent plusieurs inconvénients, l'un des inconvénients principaux est leur instabilité, sa conséquence est que les algorithmes d'apprentissage par arbres de décision ont une variance importante, qui nuit à la qualité de l'apprentissage. Des méthodes comme le Bagging (pour Bootstrap Aggregating), ou les Random Forests de Breiman permettent dans une certaine mesure de remédier à ce problème.

RF est un cas particulier du Bagging, elle est définit par les deux éléments clés suivant :

1. Les forêts aléatoires sont basées sur des ensembles d'arbres de décision.
2. Une certaine "quantité" d'aléatoire est introduite dans le processus d'induction.

Elles sont en général plus efficaces que les arbres de décision mais possède l'inconvénient d'être plus difficilement interprétables.

Chapitre 2

Fuzzy Random Forest

Introduction

Malgré la puissance de Random Forest mais les limites de ce dernier comme l'incertitude nous a pousser à améliorer ses performances , dans ce chapitre nous présentons les forets aléatoires flous (Fuzzy Random Forest) comme une extension de l'arbre de décision classique cité dans le chapitre précédent, ensuite nous présentons Fuzzy Cart.

1 Univers flou

Nous ne pouvons aborder la forêt flou sans tout d'abord présenter les bases de la théorie floue.

Dans la vie quotidienne nous appliquons implicitement la logique floue car les informations ne sont pas toujours précises. Autrement, chaque personne peut se trouver dans des situations où il utilise des informations incomplètes, elle raisonne avec elles et elle prend des décisions. Il a été nécessaire de créer une logique (Logique floue) qui admet des valeurs de vérité en dehors de l'ensemble binaire pour pouvoir tenir compte et manipuler ce genre d'information [Bar08].

1.1 Le problème

Ces imperfections peuvent être distinguées en deux classes :

– Imprécisions : désigne les connaissances qui ne sont pas perçues ou clairement

définies. Par exemple : La température de la chambre est très élevée [Der11].

- Incertitudes : désigne les connaissances dont la validité est sujette a question. Par exemple : Je crois que la température est de 30 [Der11].
- Incomplètes : du fait d'une rupture dans la transmission des données [Cor07].

Et l'intreface entre [Gér05] :

- L'information numérique (quantitatif).
- L'information symbolique (qualitatif).

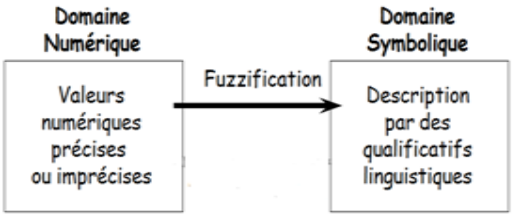

FIGURE 2.1 – Interface entre l'information numérique et symbolique

1.2 La logique floue

La logique floue est une extension de la logique classique (appelée aussi la logique booléenne), Lotfi Zadeh est le fondateur de la théorie des ensembles flous qui et défini comme : ''une collection telle que l'appartenance d'un élément quelconque a cette collection peut prendre toutes les valeurs entre 0 et 1" [Dah11].

La logique est basée sur deux concepts principaux :

- Ensembles et variables flous et opérateurs associes.
- Prise de décision a partir d'une base de règles : Si...Alors

1.3 Notion d'ensemble et sous ensemble flou

Ensemble flou

Soit X une collection d'objets notés x.

Un ensemble flou A de X est caractérisé par une fonction d'appartenance μ_A.

La valeur μ_A, pour x dans X, est comprise entre 0 et 1. Elle définit le degré d'ap-

partenance de l'objet x à l'ensemble flou A [Cor07].

Fonction d'appartenance

Il existe plusieurs formes classiques de fonctions d'appartenance floues (comme trapézoïdale, triangulaire ou gaussienne,...) qui modélisé des variables linguistiques (grand, moyen ou jeune,...). Elle permet de décrire une appartenance floue a une classe [LMAK07].

Sous ensemble flou

La logique floue permet de caractériser une appartenance graduelle a un sous ensemble, appeler sous ensemble floue [Elk10].

Dans l'approche classique :

Si μ_A est la fonction d'appartenance de l'ensemble A.

$$\forall \text{ x} \in \text{X} \quad \mu_A \text{ (x)} = 0 \quad \text{si x} \notin \text{X}$$
$$\mu_A \text{ (x)} = 1 \quad \text{si x} \in \text{X}$$

Dans l'approche floue :

- Un élément peut appartenir plus ou moins fortement a cette classe.
- Un sous-ensemble flou A d'un référentiel X est caractérisé par une fonction d'appartenance μ_A :

Si μ_A est la fonction d'appartenance de l'ensemble flou A.

$$\forall \text{ x} \in \text{X} \quad \mu_A \in [1,0]$$

Un ensemble flou est déterminer par sa fonction d'appartenance (degré d'appartenance ou valeur de vérité) [Elk10].

Exemple

Classification de personnes en 3 ensembles définit par les mots suivants :

Jeune : J, Entre deux âges : E, Agé : A

En logique classique :

En logique floue :

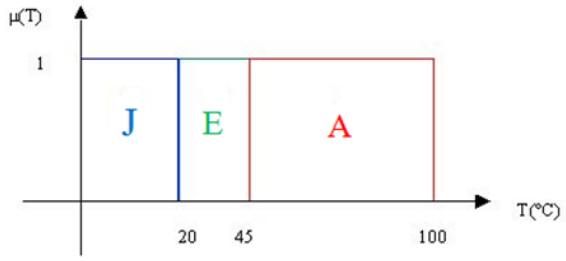

FIGURE 2.2 – La logique classique

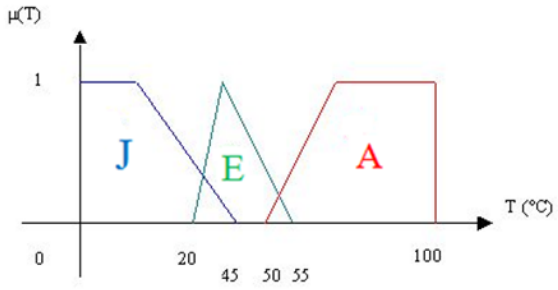

FIGURE 2.3 – La logique floue

La logique classique utilise que le 0 et le 1 ainsi les personnes sont classer en 3 catégories : jeune puis entre deux âges et en fin âgé. Au dessus nous pouvons observer la représentation graphique de trois fonctions d'appartenance Jeune, Entre deux âges et Agé qui se rapproche du raisonnement humain.

L'intérêt de l'approche "floue" vient du fait qu'un élément peut appartenir simultanément a plusieurs classe a des degrés divers [Dah11].

2 Fuzzy Random Forets (Foret Aléatoire Flou)

Les forêts aléatoires flous sont un ensemble d'arbre de décision ou on applique la logique floue.

2.1 Les forêts d'arbre de décision flou

Le principe des forêts d'arbre de décision flou

La méthode de construction d'arbre de décision est compréhensible et simple a mettre en œuvre, elle est utilisé dans différent domaine comme l'aide au diagnostic médical et la fouille des données (Data Mining) mais son problème se limite a

la mesure de discrimination quand t'il existe plus de deux classes, la solution est d'introduire une forêt d'arbres de décisions flou pour reconnaitre une classe unique.

Construire des partitions floues

L'élément important pour la prise en compte de données imprécises (l'information imparfaite) est l'existence d'une partition floue sur l'univers des valeurs d'attribut numérique – symbolique [Chr98], elle est donnée par l'expert du domaine en générale.

Pour cela, il faut un système d'apprentissage inductif autonome doté d'une méthode automatique de construction de partition floue pour gérer et traiter les données numériques en premier puis les données numériques – symboliques pour construire les arbres de décisions flous.

Les caractéristiques d'un arbre de décision flou

Les arbres de décision flous généralisent les arbres de décision classiques et son mieux adaptes pour le traitement des données numériques / symboliques dont le parcours se fait de la racine a la feuille est constitue une règle floue [AY06] pour obtenir la réponse de l'arbre.

Ils sont aussi construits avec des seuils fixés pour chaque attribut numérique qui vont être considérés comme des valeurs floues lors de l'utilisation de l'arbre générale.

Pour notre cas en utilise la méthode de Breiman existe pour construire ce seuil pendant le processus de construction de l'arbre [Chr98].

2.2 Comparaison de l'arbre décision classique avec l'arbre flou

Arbre classique (booléen)	Arbre flou
Le processus de classification suit le premier chemin valide.	Tous les chemins de l'arbre sont évalué lors du processus de classification.
Le critère de sélection pour diviser les données lors de la construction de l'arbre est pas toujours approprié.	Meilleurs pouvoir de généralisation entre l'ensembles d'apprentissage et l'ensemble de teste.
L'arbre est sensible au bruit dans l'ensemble d'apprentissage.	L'utilisation de degrés d'appartenance flous permet un traitement robuste face au bruit.
Le processus de décision dépend des valeurs seuils.	L'usage de valeurs linguistique élimine le problème des valeurs seuils.

TABLE 2.1 – La comparaison entre l'arbre de décision classique et flou [Pas04]

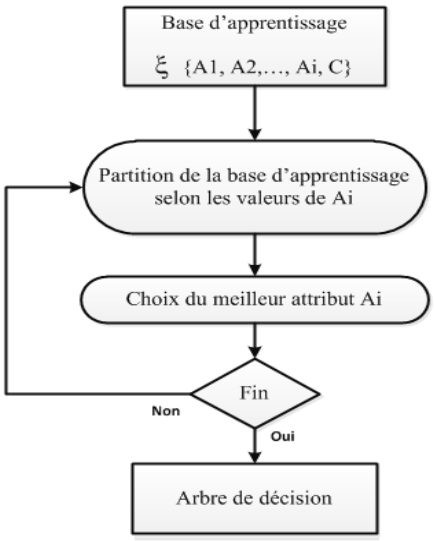

FIGURE 2.4 – Construction d'arbre de décision par la stratégie TDIDT (Top Down In – duction of Décision Tree)

ξ : La base d'apprentissage composé des exemples définit par des attributs A_i et

une classe C.

Choisir un attribut A_i pour partitionner la base d'apprentissage ξ. Un nœud est créer dans l'arbre portant un test sur la valeur A_i s'effectue généralement en décomposant la base en sous base, chaque sous base est composé d'exemples possédant une valeur identique pour A.

Si tous les exemples de la base possède la même classe, le processus de l'arbre peut s'arrêt sinon il remonte (au choix du meilleur attribut).
Dans le cadre flou on repartit les exemples de toutes les sous bases en leurs affectant un degré d'appartenance.

3 Etat de l'art

Dans les différents travaux réalisés sur le Fuzzy Random Forest, on retrouve une seule équipe de recherche du Professeur P.Bonissone qui traitent le problème des données imparfaites en utilisant des systèmes multi-classifieurs inspiré par la philosophie de Breiman basé sur des forets d'arbres de décisions floues.

1. Dans leurs papier « A Fuzzy Random Forest : Fundamental for Design and Construction » [PCGDV10] P. Bonissone et al. ont effectué des tests sur différentes bases en comparant les performances de l'algorithme FRF en appliquant des arbres de type Fuzzy ID3 avec les algorithmes de Bayes, C 4. 5, Radom forest et ADABOOST.

2. P. Bonissone et al « Combination Methods in a Fuzzy Random Forest » [PCdCGAV08] les auteurs ont utilisé deux autres méthodes de combinaisons : méthode minimum (pour chaque classe est sélectionnée la valeur minimum de toutes les feuilles atteintes dans l'arbre) et la méthode maximum (pour chaque classe est sélectionnée la valeur maximum de toutes les feuilles atteintes dans l'arbre) avec les 2 techniques FRF et FID qui est un programme qui produit un arbre de décision basé sur logique flou en permettant la gestion des données imparfaite, en particulier des étiquettes linguistiques. D'après les résultats comparatifs, les auteurs on constater que FRF : fuzzy random forest

et meilleur que FID : division of the Inspection of the forests.

3. Toujours dans la continuité de l'amélioration du principe de FRF, P. Bossinone et al dans « Weighted in a Fuzzy Random Forest » [PCG+09] ont utilisé une méthode de combinaison pondérée en utilisant une implémentation d'inférence suivant 2 stratégies :
   - S M 1 : en combinant l'information des feuilles atteint dans chaque arbre pour obtenir la décision de chaque arbre, en appliquant le même procédé de combinaison aux autres pour produire la décision globale de la forêt.
   - S M 2 : en combinant l'information des feuilles atteintes de tous les arbres pour obtenir la décision globale de la foret.

   Les résultats obtenus démontrent que l'implémentation d'inférence pondérée avec la stratégie 2 se comportent mieux qu'avec la stratégie 1.

4. Dans un papier plus globale P. Bonissone et al « A fuzzy Radom forest » [PCGDV10] ont récapitulé toutes les différentes stratégies et ont fait un bilan comparatifs entres eux. Les méthodes de combinaison utilisées sont :
   - La méthode pondérée qui dépend explicitement des données.
   - Le méthode non pondérer qui dépend des données implicites.

   Les résultats obtenus montrent de meilleurs performances (jusqu'à 65 %) pour les méthodes de combinaison pondérés par rapport aux méthodes non pondérés.

5. Dans leurs derniers papier en date l'équipe du P. Bonissone dans leurs travail intitulé « Towards the learning from low quality data in a fuzzy random forest ensemble » [CGMB11] proposent une étude comparative entre la foret floue avec des séparateurs floues non uniformes/uniformes.
   - La séparation floue non-uniforme (reconnaissance de forme de type non symétrique).
   - La séparation floue uniforme (reconnaissance de forme de type symétrique, homogène).

Les résultats obtenus montrent clairement que FRF avec les séparations flous non-uniformes donne de meilleurs résultats que les séparations floues uniformes.

6. Tout récemment le chercheur C. Marcela a presenter dans son papier « Data Mining with Ensembles of Fuzzy Decision Trees » [Mar09] une étude sur l'influence du nombre d'arbres pour la construction de la forêt d'arbre de décision flou, chacune de ces forêts a été employée pour classifier tous les exemples de l'ensemble d'essai, en utilisant 3 graphiques :
   - **Classique :** correspond au taux d'erreurs en employant un arbre de décision flou classique de type fuzzy CART .
   - **Zadeh :**
     correspond au taux obtenu en employant les t-normes de Zadeh (les opérateurs minimum et maximum) en classifiant des exemples.
   - **Lukasiewicz :**
     correspond au taux d'erreurs obtenu en employant les t-normes de Lukasiewicz en classifiant des exemples.

Contribution

Les techniques d'arbre de décision se sont révélés être interprétable et capable de traiter des applications de grande envergure, Cependant, ils sont très instables lorsque y'a de petites perturbations qui sont introduits dans l'apprentissage des données lorsqu'ils traitent avec des attributs numériques. Pour cette raison, la logique floue a été incorporée a la construction d'arbre de décision permettant l'utilisation de valeurs floues dans ce dernier.

Les FRF accroît la robustesse des arbres de décision flous, la puissance du caractère aléatoire d'augmenter la diversité des arbres dans la forêt, et la flexibilité de la logique floue peut : contrôler des données imparfaites, réduit le taux d'erreurs et met en évidence plus de robustesse et plus d'interopérabilité.

Dans le cadre de notre travail nous nous intéresserons a la construction des forets d'arbres de décision floues (les arbres de types Fuzzy CART) pour la classification de l'ensemble de la base de donnée, en utilisant la logique floue de segenou.

Conclusion

Le Fuzzy Random Forest accroit la robustesse des arbres de décision flous, il donne plus d'interprétabilité, lisibilité et la fléxibilité de la logique floue pour controler des données imparfaites.

Chapitre 3

Résultats et Intérprétations

Introduction

Dans ce chapitre nous présentons nos différentes propositions en parcourant l'état de l'art (voir chapitre 2), pour cela nous divisons ce chapitre en deux parties où chaque partie concerne une proposition. La procédure de classification des données est représentée dans le schéma suivant :

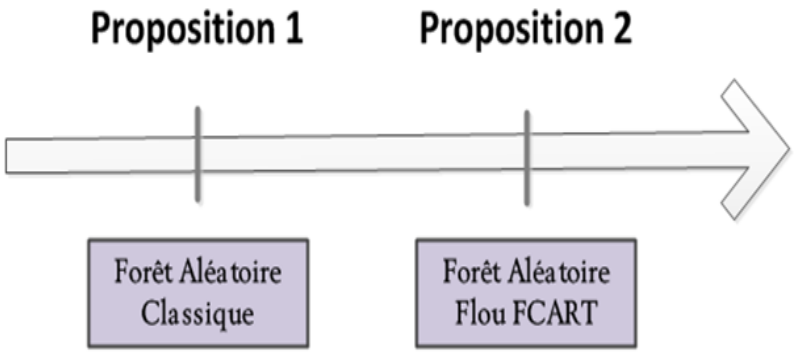

FIGURE 3.1 – Représentation de la procédure de classification dans les différentes bases.

Dans le cadre de ce mémoire, nous nous sommes intéressés à étudier la performance de l'algorithme Fuzzy Random Forest pour la classification des données. Pour cela nous avons teste deux propositions, qui concernent :

- Les forêts aléatoires classiques : Random Forest.
- L'intégration des arbres de décision flous de type Fuzzy CART pour la création des forêts aléatoires floue : Fuzzy Random Forest

1 Les bases de données

Dans notre travail les expérimentations sont effectués sur deux bases extraites de l'ensemble de données d'UCI (University Californie Irvine) [AA10].

1.1 Base de données PIMA

La base Pima Indian Diabetes (PID) [PA10a] d'Arizona est constituée de 768 femmes dont 268 sont diabétiques et 500 non diabétique. Chaque cas est formé de 9 attributs, dont le 8éme représente des facteurs de risque et le 9éme représente la classe du patient.

N°Attributs	Description de l'attribut	Moyenne	Écart type
1	Npreg : Nombre de grossesses	3.845	3.37
2	Glu : concentration du glucose plasmatique	120.895	31.973
3	BP : tension artérielle diastolique, (mm Hg)	69.105	19.356
4	Skin : épaisseur de pli de peau du triceps, (mm)	20.536	15.952
5	Insu : dose d'insuline, (mu U/ml)	79.799	115.244
6	Bmi : indice de masse corporelle,(poids en kg/(taille m)2)	31.993	7.884
7	Ped : fonction de pédigrée de diabète (l'hérédité)	0.472	0.331
8	Age : âge (Année)	33.241	11.76

TABLE 3.1 – Déscription des attributs de la base de données PIMA

N° de la Classe	Label	Nombre
1	teste négative	500 exemples
2	teste positive	268 exemples

TABLE 3.2 – Information sur les exemples de la base PIMA

1.2 Base de données BUPA

Bupa est une base de données sur les troubles hépatiques collectée par BUPA Medical Research Ltd [PA10b], elle contient 345exemples de sexe masculin 8 (200 non malades et 145 malades) définit par les 7 attributs suivants dont le dernier représente la classe :

N°Attributs	Description de l'attribut	Moyenne	Écart type
1	mcv : volume globulaire moyen	90.159	4.448
2	alkphos : alcalines phosphatase	69.87	18.348
3	SGPT : aminotransferase alanine	30.406	19.512
4	SGOT : aspartate aminotransferase	24.643	10.064
5	gammagt : gamma-glutamyl trans-peptidase	38.248	39.255
6	drinks : nombre de boissons d'une demi – pinte l'équivalent de boissons alcoolisées bu par jour	3.455	3.338

TABLE 3.3 – Déscription de la base de données BUPA

N° de la Classe	Label	Nombre
1	malade	145 exemples
2	non malade	200 exemples

TABLE 3.4 – Information sur les exemples de la base BUPA

2 Le travail effectué

Nous proposons le protocole expérimental de réparation en deux parties pour les deux bases de données :

- Les 2/3 de la base pour la phase d'apprentissage où nous utilisons le principe du Baging : Boostrap et Agrégation pour construire des sous bases qui sont égales aux nombres des arbres, sachant que ces derniers sont de types CART.
- Les données OOB (Out Of Bag) sont utilise dans la phase de validation.

- Les 1/3 restants pour la phase de test où l'erreur est estimée via un vote majoritaire.
- Le choix des paramètres des forêts aléatoires : Pour le choix du nombre d'arbres, nous proposons de faire des tests avec un nombre d'arbres variant de 10 a 500 arbres en utilisant un pas de 50 pour étudier le meilleur choix d'arbres pour les deux bases de données.

Pour chaque arbre individuel la répartition est suivi selon le principe de Breiman en utilisant un paramètre de randomisation ''Mtry" qui est égale à :$\sqrt{K}$ sachant que K= nombre de paramètres aléatoires pour chaque arbre.

2.1 Proposition 1 : Forêt Aléatoire Classique

L'algorithme de Random Forest a été introduit par Breiman en 2001, qui a définit le RF comme :

- Une famille de méthodes d'ensemble ou il permet de faire :
  1. L'agrégation d'une collection d'arbres aléatoires.
  2. Génère un arbre doublement perturbé.
- Chaque arbre de l'ensemble est généré au début d'un sous échantillon Boostrap de la base d'apprentissage complète.
- L'arbre est construit sans élagage en utilisant la méthode CART (Classification And Régression Tree) qui a la particularité d'être binaire, et utilise le critère d'indice de Gini.

Pour effectuer la répartition des 768 patients de la base PIMA et 345 patients de la base BUPA tout en gardant le principe du 2/3 pour l'apprentissage (512 pour PIMA et 230 pour BUPA) et 1/3 pour le test (256 pour PIMA et 115 pour BUPA), les résultats obtenus sont les suivants :

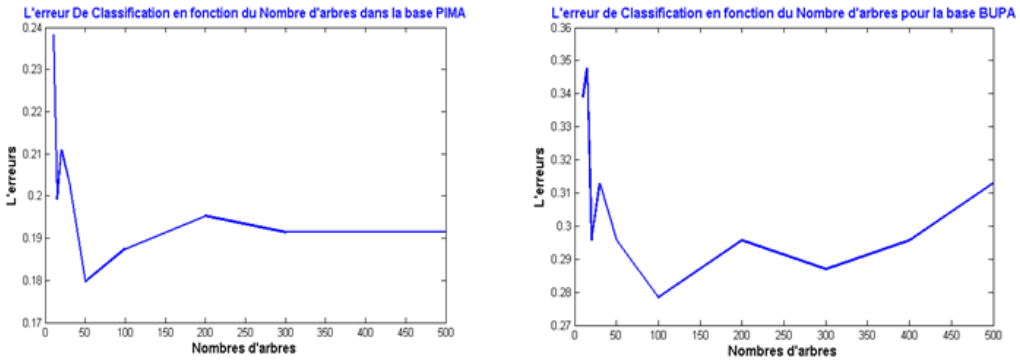

FIGURE 3.2 – Erreur de classification en fonction du nombre d'arbre pour les bases PIMA et BUPA pour RF

- La performance des forêts aléatoires (RF) sur la base PIMA a donné de meilleures performances au niveau de 50 arbres avec un taux de classification de 82% mais la stabilité est plus présente à partir de 300 arbres.
- Et pour la base BUPA le taux de classification est de 72% pour 100 arbres et en augmentant le nombre d'arbre on remarque plus au moins une stabilité à partir de 200 arbres.

L'aspect important dans le RF est la stabilité et dans les résultats obtenues on a pu constater plus au moins une stabilité notre but est d'implémenter l'arbre classique avec le flou pour voir si la stabilité sera meilleur.

2.2 Proposition 2 : FRF avec Fuzzy CART

En parcourant l'état de l'art nous avons remarqué que l'équipe du professeur Bonissone a travaillé sur le Random Forest avec des arbres de types Fuzzy ID3, d'où l'idée d'intégrer le Fuzzy CART pour la construction d'une forêt aléatoire floue.

L'arbre de décision flou FCART que nous appliquons utilise le modèle de règle de TSK (Takagi-Segenu-Kang) [SK88] flou. Où il utilise le flou juste dans la partie <Si> tandis que dans la partie <Alors> il ya des dépendances fonctionnelles :

$$\text{Si } x \in A \text{ et } y \in B \text{ Alors } z = f(x,y)$$

Dans la construction du FRF on intègre la logique floue au niveau des nœuds pour chaque arbre en utilisant l'index flou de gini qui a été introduit par Gini en in-

férence statistique, et il a été utilisé pour la construction d'arbres de décision [Chr10].

Cet index peut s'étendre pour la prise en compte de valeurs floues de la même façon que l'entropie de Shannon. L'index flou de Gini H_G (V_jU) est défini par :

$$H_G(V/U) = \sum_{i=1}^{m} p * (U_i).G_G(V/U_i), \tag{3.1}$$

avec :

$$G_G(V/U_i) = 1 - \sum_{j=1}^{m} p * (V_j/U_i)^2 \tag{3.2}$$

Sous cette forme, on peut remarquer de suite (voir équation (1.2)) que HE (V_jU) et HG (V_jU) sont construites par la même agrégation d'une fonction G (V_jU_i).

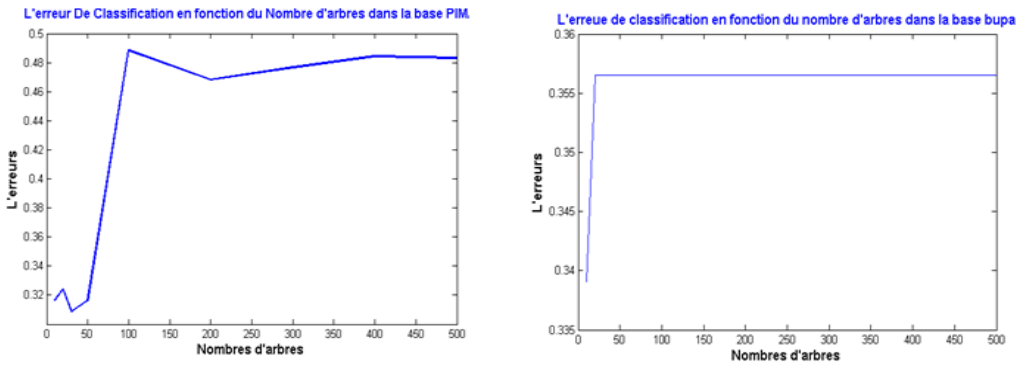

FIGURE 3.3 – Erreur de classification en fonction du nombre d'arbre pour les bases PIMA et BUPA pour FRF avec Fuzzy CART

La performance des forêts aléatoires sur la base PIMA a donné de meilleur performance au niveau de 30 arbres avec un taux de classification de 69% mais la stabilité et plus au moins présente a partir de 200 arbres, et pour la base BUPA le taux de classification est de 66% et nous remarquons que la stabilité commence à partir de 30 arbres.

La Connaissance obtenue

On distingue dans les figures suivantes les fonctions d'appartenance pour les variables de la base de données PIMA et BUPA.

– Pour la base PIMA :

Pour le cas de figure d'une foret aléatoire Floues de type Fuzzy CART avec un nombre d'arbres égal à 10 nous obtenons une base de connaissances de 60 règles de classification :

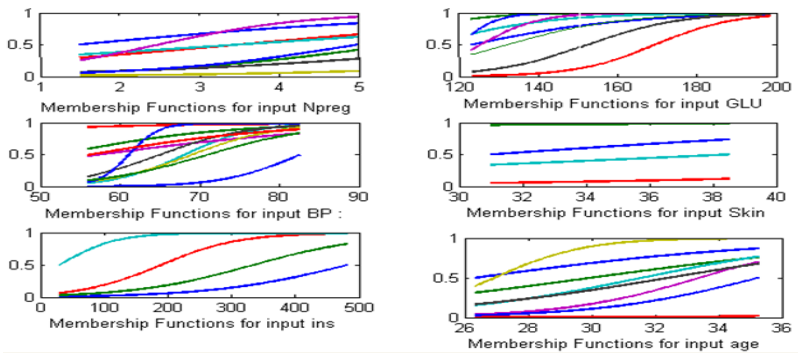

FIGURE 3.4 – Les fonctions d'appartenance de la base de données PIMA avec Fuzzy CART

On remarque clairement que la partition floue de chaque paramètre est fait de manière automatique.

– Pour la base BUPA :

Le classifieur Fuzzy CART génère une base de connaissances de 35 règles de classification :

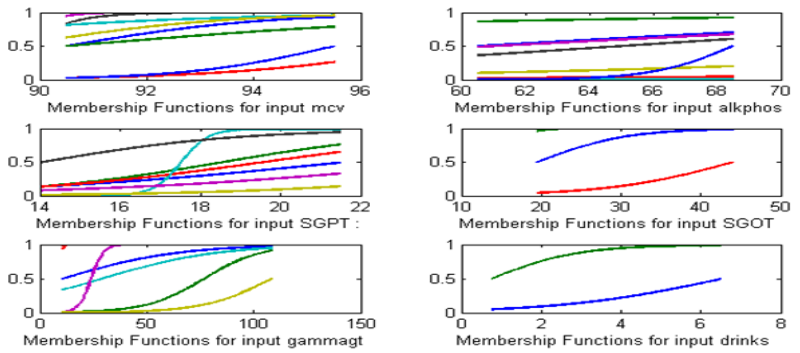

FIGURE 3.5 – Les fonctions d'appartenance de la base de données BUPA avec Fuzzy CART

Dans le même cas que la base de données PIMA , nous retrouvant une partition floue de chaque paramètre de manière automatique passant par exemple de 8 fonctions d'appartenance pour le paramètre MCV et de 2 pour le paramètre DRINKS.

L'implémentation des arbres de type Fuzzy CART dans les forêt aléatoires floues nous ont apporté la stabilité recherche dans ce travail mais en contre partie nous a fait perdre en performances avec des résultats assez médiocres avec un temps d'exécution très grand, et en interprétabilité car la connaissance obtenue est incohérente et cela peut s'expliquer par les points suivants [Yao03] :

1. Incomplétude des partitions floues : exemple deux voisins de sous-ensembles flous dans une partition floue qui ne chevauche pas.
2. Indiscernabilité des partitions floues : exemple les fonctions d'appartenance de deux sous-ensembles flous sont tellement semblables que la partition floue est indiscernable.
3. Inconsistance des règles floues : exemple Les fonctions d'appartenance perdent leurs sens prescrits physiques.
4. Trop de sous-ensembles flous.

Tous ces différents point nous ont fait réfléchir a optimiser l'apprentissage structurel des partitions floues par des méthodes de clustering dans le but d'augmenter la distinction des partitions floues et réduire le nombre de sous-ensembles flous. Tout en optimisant la conniassance.

Nous proposons l'algorithme d'optimisation qui a déjà montrer son intérêt d'application dans d'autres travaux au préalable pour l'apprentissage structurelle des paramètres floue qui est le Fuzzy C-Means (FCM).

Principe de Fuzzy C-Means (FCM)

FCM est un algorithme de clustering qui permet de voir la répartition des fonctions d'appartenances au niveau des nœuds avec le principe de regroupement pour

améliorer les performances de chaque arbre et minimiser la connaissance, il est utiliser dans la classification non supervisé (pas de classe prédéfinie).

Dans notre cas d'application le nombre de cluster va être égal à 2 car nous traitons un problème de classification binaire.

L'optimisation FCM- Fuzzy CART

L'optimisation du Fuzzy CART va se faire comme suit :

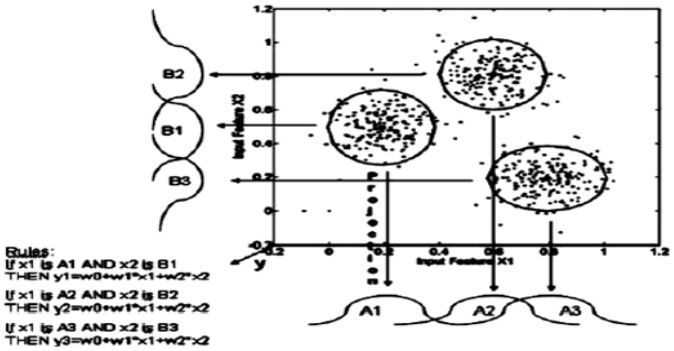

FIGURE 3.6 – Schéma représentatif de la répartition des clusters en fonctions d'appartenance avec FCM

FCM tente de partitionner les données numériques dans des clusters. L'appartenance d'un point de données à un cluster spécifique est exprimée par la valeur d'appartenance de ce point à ce cluster. La valeur d'appartenance est calculée par la minimisation d'une fonction objective de FCM, qui recherche l'appartenance ressortant le moins d'erreur.

Nous allons effectuer la même répartition des bases de données faite dans RF et FRF avec Fuzzy CART, les résultats obtenus sont les suivants :

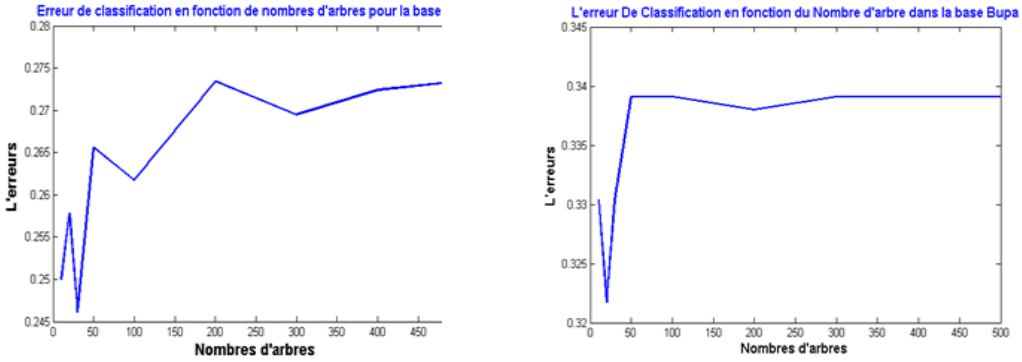

FIGURE 3.7 – Erreur de classification en fonction du nombre d'arbre pour les bases PIMA et BUPA pour FRF avec Fuzzy C-Means

Les performances des forêts aléatoires sur la base PIMA a donner de meilleur performance a niveau de 30 arbres avec un taux de classification de 76% mais la stabilité et plus au moins présente a partir de 300 arbres, et pour la base BUPA le taux de classification est de 68% et nous remarquons que la stabilité commence a partir de 20 arbres.

La Connaissance obtenue

Le classifieur Fuzzy FCM génère une base de connaissances de 2 règles de classification pour les deux bases :

– Pour la base PIMA :

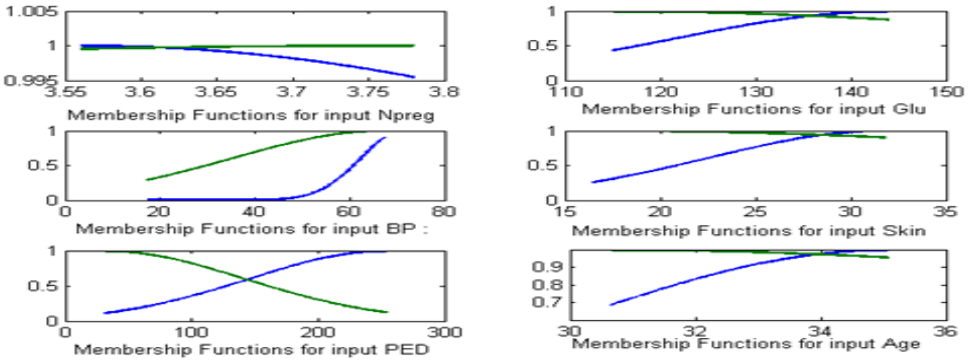

FIGURE 3.8 – Les fonctions d'appartenance de la base de données PIMA avec Fuzzy C-Means

La connaissance extraite du modele FCM

Règle 1 : If (Npreg is **petit**) and (Glu is **grand**) and (Bp is **grand**) and (Skin is **grand**) and (PED is **grand**) and (Age is **grand**) then (class is **malade**).

Règle 2 : If(Npreg is **grand**) and (Glu is **petit**) and (Bp is **petit**) and (Skin is **petit**) and (PED is **petit**) and (Age is **petit**) then (class is **non malade**).

– Pour la base BUPA :

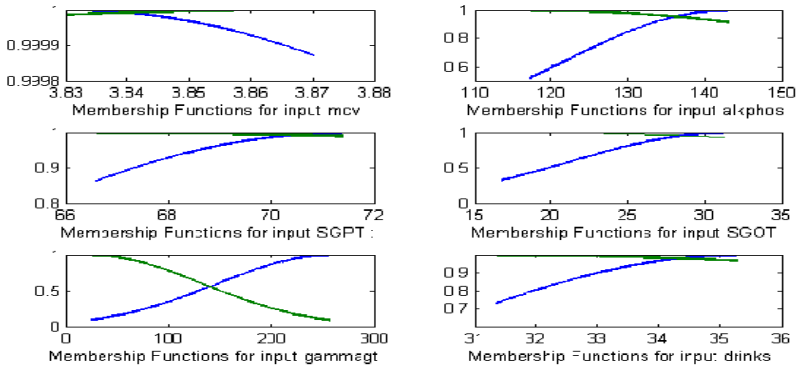

FIGURE 3.9 – Les fonctions d'appartenance de la base de données BUPA avec Fuzzy C-Means

La connaissance extraite du modele FCM

Règle 1 : If (mcv is **grand**) and (alkphos is **grand**) and (SGPT is **grand**) and (SGOT is **petit**) and (gammagt is **grand**) and (drinks is **grand**) then (class is **malade**).

Règle 2 : If (mcv is **petit**) and (alkphos is **petit**) and (SGPT is **petit**) and (SGOT is **petit**) and (gammagt is **petit**) and (drinks is **petit**) then (calss is **non malade**).

Dans cette partie, chaque attribut d'entrée a deux fonctions d'appartenance. On a remarqué que l'application de Fuzzy C-Means a réduit considérablement le nombre de règles qui a été minimisé de 60 règles à 2 pour PIMA et de 35 à 2 pour BUPA,

ce qui diminue la complexité de la base de connaissances de manière significative.

3 Synthèse des différentes propositions

Pour des fins de comparaison avec les différents tests et résultats obtenues au par avant nous avons voulu compléter cette synthèse en implémentant le Fuzzy ID3 (Id3 : induction of decision tree) utiliser dans les travaux de l'état de l'art ou l'intégration du flou se fait au niveau de la racine (dans les imput), pour reconfirmé les résultats de ce dernier qui a donner de meilleur résultat la preuve, voici les résultats avec un nombre d'arbre fixé a 10 :

	Fuzzy CART	Fuzzy FCM	Fuzzy ID3
TC % PIMA	68,36%	75%	88,67%
TC % BUPA	66,09%	68%	86,37%

TABLE 3.5 – Les résultats

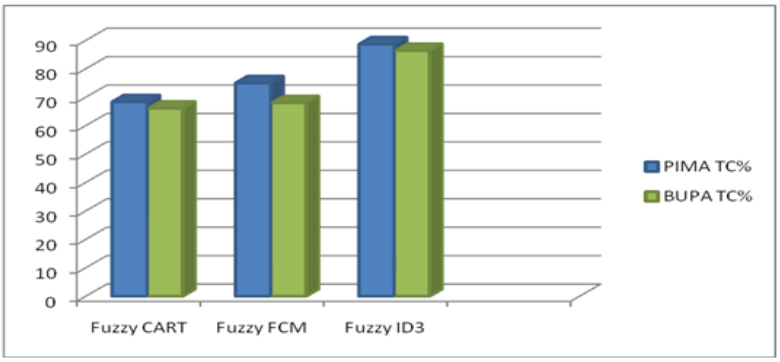

FIGURE 3.10 – Histogramme des performances des différentes techniques

Nous remarquons que Fuzzy ID3 a bénéficié des meilleurs résultats obtenus pour les performances de chaque classifieurs dans les deux bases de données, il a augmenté visiblement ses résultats par rapport aux autres classifieurs. Ce qui permet de confirmé la contribution de l'état de l'art.

Conclusion

Cette étude présente un modèle de classification de forêts aléatoires flous, pour cela nous avons utilisé deux critères pour évaluer la méthode proposée. Le premier est la stabilité des performances du classifieur, le second est le bon taux de classification obtenue par l'erreur.

L'algorithme de Fuzzy C-Means (FCM) a permis d'optimiser l'algorithme de Fuzzy CART en réduisant son temps d'exécution mais tout en gardant une très bonne stabilité et un très bon taux de classification.

Avec l'hybridation de ces méthodes une extraction de connaissances obtenues de fuzzy inférence système a pu être faite passant de 60 règles sur le Fuzzy CART à 2 avec Fuzzy FCM ou la connaissance est plus fiable, précise et suffisamment simple pour être comprise, le tout en améliorant les performances avec un bon taux de classification pour les deux bases de données.

Conclusion générale

Dans ce travail nous avons présenté une méthode de classification des forêts aléatoires qui est un cas particulier de méthodes d'ensemble (bagging) apres hybridation de ce dernier avec le flou.

Les arbres de décision flous présentent un avantage majeur dans la classification de par leur simplicité, et leur facilité d'interprétation. L'induction des règles de décision à partir de l'arbre induit représente l'un de ses avantages principaux. Notons que la méthode que nous présentons dans ce mémoire répond à un besoin vital dans le domaine médical et offre aux médecines une base de connaissance explicite (sous forme de règles) acquise d'une base de données médicale. L'expert aura la possibilité d'accepter les règles, de les modifier, de les supprimer ou d'ajuter d'autres.

Nous avons réussi à implémenter un classifieur basé sur l'arbre de décision flou de type fuzzy cart, les résultats obtenus sont non satisfaisants, en terme de temps d'exécution et taux de classification, d'où l'initiative d'améliorer cette approche par algorithme d'optimisation Fuzzy C means. Il permet une meilleure répartition des données et ainsi qu'une régularisation des contraintes qui s'appliquent sur les paramètres des fonctions d'appartenance floues. Cet algorithme réduit le nombre de sous-ensembles flous et minimise le nombre de règles pour une connaissance ciblée. Delà donc automatiser et optimiser la structure et les paramètres des fonctions d'appartenance.

Nous pouvons dire à partir de ces résultats obtenus, que les arbres de décision flous constituent une technique importante dans tout système pour la modélisation de la connaissance dans l'aide au diagnostic médical en général.

Nous pouvons dire à partir de ces résultats obtenus, que les arbres de décision flous constituent une technique importante dans tout système pour la modélisation de la connaissance dans l'aide au diagnostic médical en général.

Dans les perspectives de ce travail, nous nous intéressons à la connaissance extraite en modélisant une hiérarchie d'arbres qui ont le plus de contribution dans les performances et la bonne classification des données médicales.

Bibliographie

[AA10] A.Frank and A. Asuncion. UCI machine learning repository, 2010.

[Aur06] VESIN Aurélien. *Arbre de décision*, 2006.

[AY06] Chikh Mohammed Amine and Glorennec Pierre Yves. Application des arbres de décision flous á la reconnaissance des bvps. *GBM U. Tlemcen Algérie, et INSA Rennes France, seconde édition des Journées d'Etude algéro-francaises en Imagerie Médicale*, USTHB 21-22 Novembre 2006.

[Bar08] Rachid Baroudi. Une approche d'estimation de la criticité des agents basée sur la classification floue. In *CIIA*, 2008.

[Beh09] Omar Behadada. Construction d'un système d'aide à la décision médicale pour la détection des arythmies cardiaques à l'aide d'arbres de décision flous. In *CIIA*, 2009.

[Ber09] Simon Bernard. Forêts aléatoires : De l'analyse des mécanismes de fonctionnement à la construction dynamique. December 2009.

[BHAO09] Simon Bernard, Laurent Heutte, Sébastien Adam, and Emilie Oliveira. On the selection of decision trees in random forests. In *IJCNN*, pages 302–307, 2009.

[Bre01] Leo Breiman. Random forests. *Machine Learning*, 45(1) :5–32, 2001.

[CGMB11] José Manuel Cadenas, M. Carmen Garrido, Raquel Martinez, and Piero P. Bonissone. Towards the learning from low quality data in a fuzzy random forest ensemble. In *FUZZ-IEEE*, pages 2897–2904, 2011.

[Chr98] Marsala Christophe. Application of fuzzy rule induction to data mining. In *FQAS*, 1998.

[Chr10] Marsala Christophe. Apprentissage artificiel et raisonnement flou. In *FUZZ-IEEE*, 30 novembre 2010.

[Cor07] Laurence Cornez. Discrimination automatique à base de connaissances expertes d'événements sismiques. In *CIIA*, 2007.

[Dah11] Mostafa El Habib Daho. optimization du systeme nefclass. Master's thesis, Université Aboubekr Belkaid, Juin 2011.

[Der11] Franck Dernoncourt. Fuzzy logic : between human reasoning and artificial intelligence. Master's thesis, ENS Ulm, January 2011.

[Elk10] Sabeur Elkosantini. *Introduction a la logique floue : Les concepts fondamentaux et applications*, 2010.

[Gen10] Robin Genuer. Forêts aléatoires : aspects théoriques, sélection de variables et applications. 2010.

[Gér05] Dray Gérard. Extraction de connaissances à partir de données et fouille de données. In *CORIA*, 2005.

[LMAK07] H. LAANAYA, A. MARTIN, D. ABOUTAJDINE, and A. KHENCHAF. Régression floue et crédibiliste par svm pour la classification des images sonar. In *Extraction et Gestion des Connaissances (EGC)*, pages 21–32, Namur, Belgique, 24-26 January 2007.

[Mar09] Christophe Marsala. Data mining with ensembles of fuzzy decision trees. In *CIDM*, pages 348–354, 2009.

[PA10a] A. Frank Pima and A. Asuncion. Pima indians diabetes dataset. In *UCI Machine Learning Repository, University of California, Irvine*, 2010.

[PA10b] A. Frank Pupa and A. Asuncion. Bupa , liver-disorder. In *UCI Machine Learning Repository, University of California, Irvine*, 2010.

[Pas04] Garcia Pascal. *Utilisation d'arbres de décision flous.* PhD thesis, l'INSA de Rennes, July 2004.

[PCdCGAV08] Piero P.Bonissone, José Manuel Cadenas, Maria del Carmen Garrido, and Ramon A.Diaz-Valladares. Combination methods in a fuzzy random forest. In *SMC*, pages 1794–1799, 2008.

[PCG+09] Piero P.Bonissone, José Manuel Cadenas, M.Carmen Garrido, Ramon A.Diaz-Valladares, and Raquel Martinez. Weighted decisions in a fuzzy random forest. In *IFSA/EUSFLAT Conf.*, pages 1553–1558, 2009.

[PCGDV10] Piero P.Bonissone, José Manuel Cadenas, M. Carmen Garrido, and R. Andrés Diaz-Valladares. A fuzzy random forest. *Int. J. Approx. Reasoning*, 51(7) :729–747, 2010.

[Sco12] Erwan Scornet. Apprentissage et forêts aléatoires. *Knowl-Based Syst.*, 44 :48–56, 2011-2012.

[SK88] M. Sugeno and G.T. Kang. Structure identification of fuzzy model. In *Fuzzy Sets Syst*, 1988.

[Sté11] Caron Stéphane. Une introduction aux arbres de décision. In *CHES*, 2011.

[Yao03] Jin Yaochu. Advanced Fuzzy Systems Design and Application. 2003

Résumé

Nous traitons dans ce mémoire l'extraction de la connaissance à partir des données, en utilisant les forets aléatoires floues qui combinent la robustesse des arbres de décision, la puissance du caractère aléatoire qui augmente la diversité des arbres dans la forêt, et la flexibilité de la logique floue. Ils ont la spécificité de contrôler des données imparfaites, de réduire le taux d'erreurs et de mettre en évidence plus de robustesse et plus d'interprétabilité.

Dans le cadre de notre travail nous nous intéresserons a la construction une forêt d'arbres de décision floues (de types Fuzzy CART) pour la classification de données médicales, nous optimisons ensuite ces arbres avec l'algorithme Fuzzy C-Mean qui nous permettra une meilleure répartition des données et ainsi qu'une régularisation des contraintes qui s'appliquent sur les paramètres des fonctions d'appartenance floues. Cet algorithme réduit le nombre de sous-ensembles flous et minimise le nombre de règles pour une connaissance ciblée.

Abstract

We deal in this paper the extraction of knowledge from data, using fuzzy random bits that combine the robustness of decision trees ; the power of randomness increases the diversity of trees in the forest, and flexibility fuzzy logic. They control the specificity of imperfect data, reduce the error rate and highlight more robustness and interpretability.

In our work we focus on the construction of a fuzzy decision tree forest (types Fuzzy CART) for the classification of medical data, we then optimize these trees with Fuzzy C-Mean algorithm that will allow us to better distribution of data and a regularization of the constraints that apply to the parameters of fuzzy membership functions. This algorithm reduced the number of fuzzy sets and minimizes the number of rules for a specific knowledge.