Construction d'un système d'aide à la décision médicale pour la détection des arythmies cardiaques à l'aide d'arbres de décision flous

Omar Behadada doctorant chercheur Laboratoire de génie biomédical
Département d'électronique Biomédicale
E-mail : O_behadada@mail.univ-tlemcen.dz
M.A Chikh maître de conférences Département d'informatique
E-mail : mea_chikh@mail.univ-tlemcen.dz
Laboratoire de génie biomédical, Département d'électronique Biomédicale
Faculté des Sciences de L'ingénieur ;
Ammar Mohammed magister, Laboratoire de génie biomédical
Département d'électronique Biomédicale
E-mail : amm1222k@yahoo.fr

Université ABOUBEKR BELKAÏD
Tel: 213 43 28 56 89, Fax: 213 43 28 56 86

Résumé. L'extraction de connaissances est un processus interactif et itératif d'analyse d'un grand ensemble de données brutes afin d'en extraire des connaissances exploitables, et où l'utilisateur-analyste (cardiologue) joue un rôle central. Dans la perspective de conception de systèmes d'extraction de connaissances et de classification à partir d'une base de données cardiologiques, nous présentons une méthode basée sur les arbres de décision flous. La première partie présente la problématique du choix des caractéristiques d'un battement cardiaque. Ensuite nous appliquons l'arbre de décision flou pour la classification de quelques anomalies cardiaques. Dans la troisième partie nous montrons l'intérêt des règles de décision extraites dans l'interprétabilité des résultats de la classification.

.

Mots clés : Extraction de connaissances, arbre de décision flou, base de données cardiologiques, sélection des attributs, règles de décision.

1 Introduction

De nos jours, les nouvelles technologies de l'information produisent de nouvelles approches méthodologiques tentant d'en extraire non seulement une information valide et fiable, mais plus généralement des connaissances permettant d'étayer la décision. Deux grands types d'approches peuvent être distingués aux analyses des données : l'une est décisionnelle et s'attache le plus souvent à la modélisation, l'autre est exploratoire et a pour objectif de synthétiser un ensemble d'informations plus ou moins hétérogènes (l'approche est alors essentiellement descriptive).

Les nouveaux outils d'extraction des connaissances à partir des données (ECD) s'inscrivent clairement dans le versant exploratoire des études statistiques mais s'enracinent également dans ce second type de logique. Certaines de ces méthodes sont récentes, le concept d'ECD apparaît pour la première fois en 1989.

Dans ce présent article nous avons appliqué la méthode de l'arbre de décision flou pour extraire des règles de décision afin de construire un système de classification de quelques anomalies cardiaques.

Un arbre de décision est un outil d'aide à la décision et à l'exploration de données. Il permet de modéliser simplement, graphiquement et rapidement un phénomène mesuré plus ou moins complexe. Sa lisibilité, sa rapidité d'exécution et le peu d'hypothèses nécessaires à priori par contre La logique floue a été proposée par Zadeh [Zadeh, 1965] pour modéliser l'information et pour se rendre compte du caractère vague des connaissances que nous, les humains, manipulons au quotidien, le couplage de la logique floue avec les arbres de décision rend les règles extraites plus linguistiques et plus explicites.

2. Prétraitement de la base de données :

Nous avons collecté les données cardiaques des différents battements pour les différents enregistrements à partir de la base de données MIT-BIH avec les anomalies cardiaques ciblés (tableau 1).

Tab.1 : Les différents enregistrements choisis avec le nombre d'exemples sélectionnés

Enregistrement	Nbre 'N'	Nbre 'V'	Nbre 'R'	Nbre 'L'
100	62	0	0	0
101	5	0	0	0
103	58	0	0	0
105	10	0	0	0
106	27	34	0	0
109	0	0	0	104
111	0	0	0	41
113	6	0	0	0
115	10	0	0	0
116	45	0	0	0
118	0	0	12	0
119	50	34	0	0
122	5	0	0	0
123	5	0	0	0
124	0	0	33	0
200	0	25	0	0
203	0	15	0	0
207	0	0	0	40
208	0	152	0	0
212	5	0	26	0
214	0	50	0	50
215	103	0	0	0

2.1. Caractérisation du battement cardiaque

Un battement cardiaque est caractérisé par une succession d'onde de nature électrique (électrocardiogramme ECG), il présente un grand intérêt diagnostic.

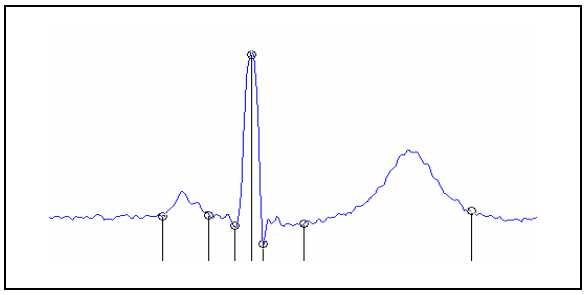

Figure 1 : Différentes ondes à détecter par IMPE

Caractérisation du battement cardiaque par des descripteurs pertinents est indispensable lors de la conception et l'implémentation de tout modèle de reconnaissance d'une anomalie cardiaque. Il convient de remarquer que de nombreuses approches citées dans la littérature ont porté sur la difficulté que représentent la mesure et le choix des paramètres pertinents du signal ECG et leur classification. On peut citer les travaux menés par plusieurs chercheurs, pour la réduction du complexe QRS pour chaque battement cardiaque. Acharya et al. [Acharya 2004][9] ont utilisé l'entropie spectrale, les déviations standards et la mesure de l'exposant de Lyapunov de la variation rythmique, pour la classification neuronale des arythmies cardiaques. Zhou [Zhou 2003][10] a appliqué la méthode de l'ACP pour réduire la taille du complexe QRS. Le vecteur réduit est présenté ensuite à l'entrée d'un réseau neuronal, pour la détection du battement ventriculaire prématuré. Lagerholm et Person [Lagerholm 2000][2] ont Caractérisé chaque cycle cardiaque par son intervalle RR et les coefficients résultant de la décomposition du complexe QRS en fonctions d'Hermite de base.

Dans notre travail nous avons conçu notre propre interface de mesures qui permet de localisé les différents ondes.

2.2. Sélection des attributs :

La sélection de nos attributs est liée avec la connaissance médicale cardiologique (tableau 2).

Tab.2: Les différents descripteurs

Paramètres	**Signification**
Durée P	Largeur de l'onde P
Intervalle PR	Du début de l'onde P jusqu'au début du QRS
Complexe QRS	Début de l'onde Q jusqu'à la fin de l'onde S

Segment ST	De la fin de l'onde S ou R jusqu'au début de l'onde T
Intervalle QT	Du début du QRS jusqu'à la fin de l'onde T
RRp	La distance entre le pic R du présent battement et le pic R du battement précédent.
RRs	Entre le pic R du présent battement et le pic R du battement suivant.
RDI (retard de la déflexion intrinsecoïde)	Du début du QRS jusqu'au sommet de la dernière positivité de l'onde R.
Durée battement	Début de l'onde P jusqu'à la fin de l'onde T.
RRs \ RRp	Le rapport RRs\RRp

2.3. Exploration de la base de données :

Une exploration de la base de données permet de voir clairement la corrélation entre les attributs et la classe ciblé.

2.3.1. Représentation bidimensionnelle :

Nous avons testé séparément les attributs en fonction de leurs classes sachant qu'ils sont codés (0,1,2 et 3).

0 :c'est le cas normal (N).

1 :c'est le cas extrasystole ventriculaire (V).

2 : c'est les cas bloc de branche droit (R).

3 : c'est les cas bloc de branche gauche (L).

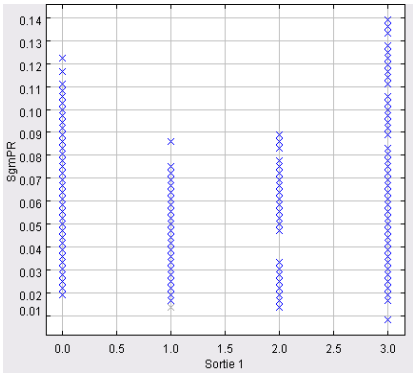

Figure 3 -a : Une représentation de segment P-R en fonction des classes

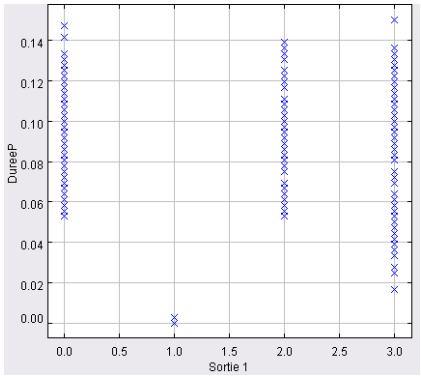

Figure 3-b : Une représentation de durée P en fonction des classes

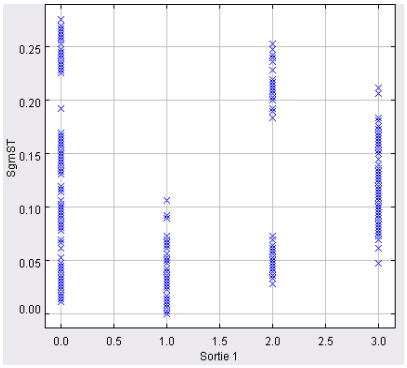

Figure 4-a : Une représentation de segment ST en fonction des classes

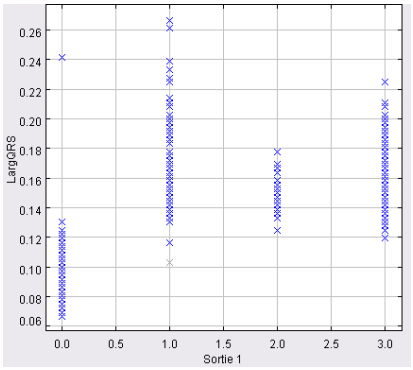

Figure 4 -b : Une représentation de largeur QRS en fonction des classes

Commentaires :

Nous voyons clairement que la durée de P dans la (figure 3 (a)) a parfaitement la même plage de variation pour les 3 classes (N, R, et L) et une variation presque nulle pour la classe V ce qui est très logique du point de vue physiologique (dans le cas d'une extrasystole ventriculaire (V), l'onde P est absente).Par contre le segment PR varie d'une manière très similaire pour les quatre clases (figure 3 (b)), donc son effet est négligeable pour l'entrée du classifieur. Le complexe QRS varie différemment dans le cas normal(N) et dans les cas pathologiques (V, R, et L) (figure 4 (a)). Les autres paramètres varient différemment d'une classe à une autre, Ce qui peut être très utile dans le renforcement des paramètres du classifieur. Nous déduisons qu'un seul paramètre seul ne peut pas être discriminant pour les différentes classes.

2.3.2. Représentation tridimensionnelle

Dans cette partie, nous avons testé l'effet de deux paramètres ensemble à la fois sur les 4 classes ciblées. La couleur rouge (+) représente le cas normal .La couleur verte (+) représente le cas extrasystole ventriculaire. La couleur bleue (+) représente le cas bloc de branche droit. La couleur violette (+) représente le cas bloc de branche gauche.

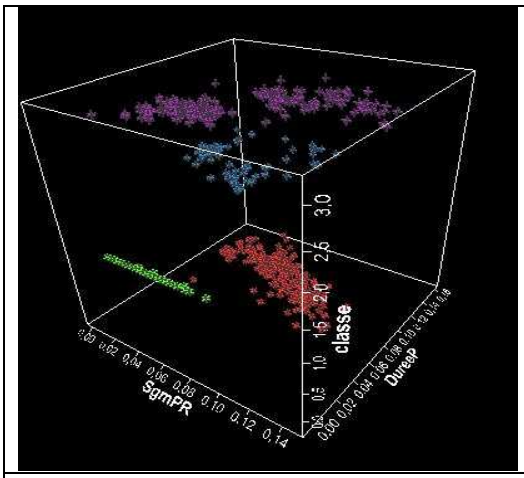

Figure 5-a : Représentation 3D de la durée P et le segment PR en fonction des classes

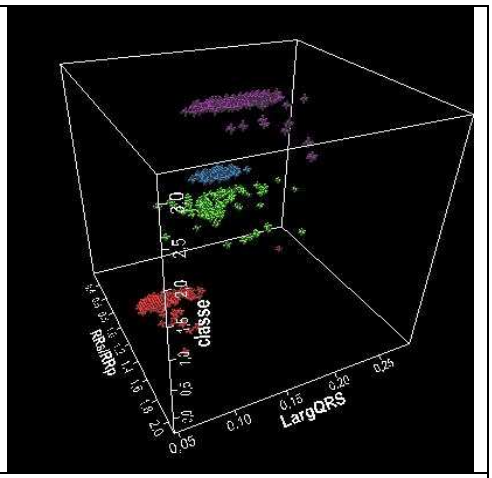

Figure 5-b : Représentation 3D de la largeur du QRS et le RRs/RRp en fonction des classes

Commentaire :

Nous avons déjà remarqué, la limite d'un seul paramètre à discriminer entre les différentes classes dans la 1ère expérimentation, ce qui nous a poussé à voir l'intérêt d'augmenter le nombre des attributs. Et nous remarquons clairement dans la (figure 5(a)) que seul le nuage de points pour le cas durée P nulle (couleur verte) se distingue des autres, ce qui correspond à la classe extrasystole ventriculaire, par contre pour le dernier couple (largeur QRS, RRs/RRp) (figure 5(b)) nous distinguons une bonne discrimination entre les pathologies, les nuages de points des différentes classes sont nettement séparés. Nous pouvons tirer de ces expérimentations menées sur les différents paramètres du vecteur d'entrée d'un classifieur des arythmies cardiaques les points suivants :

1- Les performances de tout modèle de classification, indépendamment de la technique utilisée, dépend énormément du vecteur d'entrée.

2- Une représentation géométrique des paramètres d'entrée peuvent être intéressante sur le plan visuel et sur le choix final de ces paramètres. Mais néanmoins une visualisation de plus de deux paramètres devient difficile à représenter.

3- Des paramètres comme : RRs, RRs/RRp, Durée P et largQRS sont très pertinents pour la reconnaissance des pathologies comme V, R et L. Une bonne mesure des différents paramètres reste décisive pour la suite de la classification.

3. Rappel sur l'induction d'un arbre de décision flou(ADDF)

Considérons le problème général suivant. Soit f(x1,…,x n) y=f(x1,…,x n) une relation d'entrée/sortie inconnue, les entrées x1 à x n étant des variables explicatives potentielles et y la sortie. Parmi les entrées, certaines sont importantes, d'autres redondantes, d'autres enfin inutiles. L'induction d'un ADDF consiste à :

Hiérarchiser les variables d'entrée en fonction de leur importance.

Évaluer l'utilité de prendre en compte ou non certaines entrées, soit en ligne, soit hors ligne par élagage.

Il faut, pour cela, un ensemble d'apprentissage,

E = { (xi, yi) ;x i = (x i1,…,x iM),yi $\in$ R , pour i = 1 à P } , où les xij , pour j = 1 à M,

Sont des variables linguistiques possédant m j fonctions d'appartenance, (Ajk)mkj=1.

Les fonctions d'appartenance forment des partitions floues triangulaires sur les domaines d'entrée. Sans perte de généralité, la T-norme utilisée est le produit :

ET(x, y) = x * y. ces choix apportent la propriété suivante :

$$\sum_F \alpha_F(x) \equiv 1 \qquad (1)$$

Et l'équation (1) devient :

$$\text{ADDF(x)} = \sum_F \alpha_F(x) c_F \qquad (2)$$

La construction automatique d'un arbre nécessite des mesures comme l'entropie et le gain d'information.

Dans un problème de classification, dans le cas idéal, les vecteurs d'apprentissage associés à un nœud terminal (une feuille) appartiennent à la même classe.

On dit alors que la feuille est "pure". Ce n'est évidemment pas toujours possible et l'objectif de l'induction vise à créer des feuilles avec un degré de mélange minimum.

Le partitionnement est réalisé à chaque nœud par des tests portant sur une variable : il faut donc choisir le meilleur test, l'entropie permet de faire le bon choix de la classe.

- la notion de représentation de la classe k au nœud N pour une modalité Aj de la variable traitée, joue un rôle central. Elle est définie par :

$$\text{r (k, j, N)} = \sum_{i=1}^{P} \mu_k(x_i) \wedge \mu_{Aj}(x_{ij}) \wedge \alpha_N(x_i) \qquad (3)$$

On en déduit les paramètres Pk et wj utilisés pour le calcul du gain:

$$\text{Pk} = \frac{\sum_j r(k, j, N)}{\sum_k \sum_j r(k, j, N)} = \frac{\sum_j r(k, j, N)}{\sum_{i=1}^{P} \alpha_N(X_i)} \qquad (4)$$

$$\text{wj} = \frac{\sum_k r(k, j, N)}{\sum_k \sum_j r(k, j, N)} = \frac{\sum_k r(k, j, N)}{\sum_{i=1}^{P} \alpha_N(x_i)} \qquad (5)$$

Le gain d'information apporté par un attribut X au nœud N est :

G(X, N)=I(N)-Info(X,N) (6)

Avec

I(N)=$-\sum_k P_k \log P_k$ Information au nœud N (entropie) (7)

Info(X,N)=$\sum_j w_j I(X_j)$ Information apportée par X au nœud N (8)

4. Construction du système d'inférence flou (SIF)

Le deuxième point important dans l'extraction des règles et surtout pour le cas d'induction des règles flous,est le choix des modalités floues car ce dernier représente l'aspect linguistique dans le raisonnement flou.

Une modalité floue ou le sous-ensemble flous avec les attributs forme une règle de la forme :

« Si attribut k est SEF i et attribut k' est SEF i '' alors classe C »

Nous concluons qu'un choix rigoureux des modalités floues doit être fortement exigé afin de pouvoir parler d'une induction des règles réussite.

Dans notre approche nous avons choisi les modalités floues selon la nature de variation des données en gardant les paramètres d'entrée présentés dans le tableau suivant :

Tous les sous ensemble flous (SEF) sont initialisés après analyse sauf pour le cas de l'arbre Addf1 ou nous les avons initialisés manuellement (avec l'avis du cardiologue).

Tab.3. attributs avec le nombre de SEF

Attributes	Addf1	Addf2	Addf3
DurP	3	3	3
Seg PR	2	3	3
LargQRS	3	3	3
SegST	2	3	inactive
InterQT	3	3	3
RR p	2	3	3
RR s	2	3	3
RDI	2	3	3
Durée bat	2	3	inactive
RRs /RRp	2	3	3

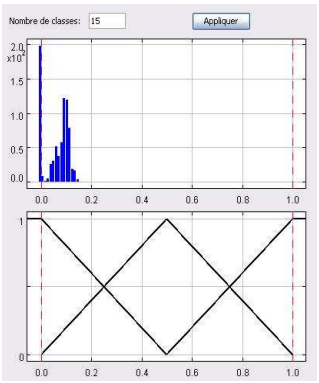

Figure 5-a : Histogramme de la durée avec la partions floue

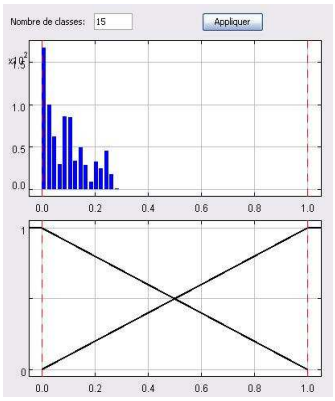

Figure 5-b : Histogramme de segment St avec la partions floue

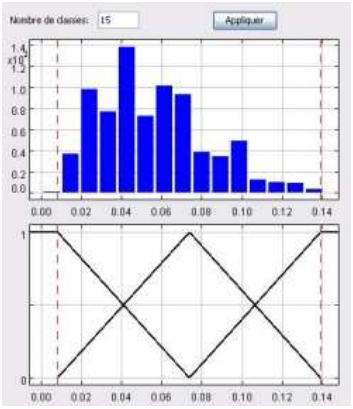

Figure 6-a : Histogramme du segment PR avec la partions floue

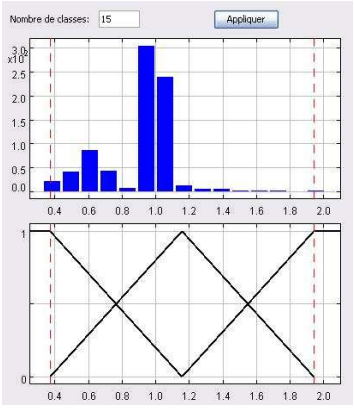

Figure 6-b : Histogramme du RRs/RRp avec la partions floue.

Les figures 5 et 6 montrent clairement la relation entre la distribution des données, différents attributs et le choix de ses modalités floues.

Nous remarquons que le choix de trois modalités floues pour attributs segment PR est très bien et uniformément réparti par contre le choix de 3 modalités floues pour l'attribut durée P avec des modalités initialisées manuellement n'est pas représentatif.

5. Conception du modèle de classification :

En vue de bien voir l'effet de nos choix déjà faits et afin de pouvoir extraire une connaissance depuis notre base de données notre approche d'extraction de connaissances c'est utiliser l'apprentissage supervisé par arbre de décision flou en construisant un système capable de reconnaître des arythmies cardiaques avec une base de connaissances comme référence. Cet objectif nous ramène à évaluer la qualité des classifieurs conçus et exploiter la connaissance induite après apprentissage de l'arbre (classifieur) qui donne les meilleur résultats.la boite à outils FisPro[10] logiciel dédié au système d'inférence flou.

5.1. Critères de performances :

Paramètres	Addf1	Addf2	Addf3
True positive	203	206	206
True negative	58	341	307
False positive	470	84	100
False negative	3	0	0
Sensitivity (se)	98,543	100	100
Specificity (SP)	10,984	80,023	75,429
Rate FP	89,015	19,764	24,570
CC	30,163	71,103	67,320

5.2. Résultats obtenus :

Nous constatons que le meilleur taux de reconnaissance est obtenu par le classifieur ADDF2 ce qui nous confirme le bon choix des modalités floues et la sélection des attributs.

Une sensibilité de 100% et avec un faut négatif de 0 nous montre la reconnaissance de l'anomalie ESV a 100%.

Par contre le classifieur ADDF1 avec un taux de reconnaissance de 30,16% à cause de la mauvaise initialisation des points modaux.

5.3. Analyse des règles de classification à partir d'un arbre de décision flou:

L'avantage principal d'un arbre de décision flou c'est l'interprétabilité des résultats et aussi leur capacité à extraire la connaissance d'une base exemples ce qui constitue un intérêt majeur dans un système d'aide au diagnostic [12].

Cette connaissance se traduit par un ensemble de règles sous forme de :

« Si A et SEF1 et B est SEF2 et…alors c'est C1 »

Discussion :

A fin de mieux évaluer la qualité de notre connaissance induite par l'arbre de décision flou nous allons analyser les règles du deuxième arbre qui a donné des meilleurs résultats par rapport aux autres avec une bonne classification et des règles très significatives et très crédibles et conformes avec l'expertise humaine.

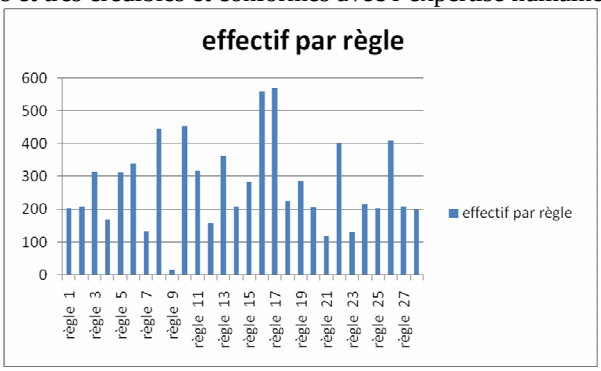

Figure 7 : histogramme présentant les règles en fonction du nombre d'exemples l'activant

En regardant l'histogramme de la figure 4-20 nous voyons clairement les règles principales qui sont très activées et d'autres moins, avec un nombre d'exemples différents, la règle 9 est activée avec peu d'exemples (seulement 17 exemples).

La règle 7 « Si durée P petite et QRS grand et RRs/RRp petite alors ESV »est activée avec 134 exemples, cette règle est vérifiée physiologiquement.

Le résultat de notre application se présente sur la table des règles où on remarque les attributs avec leur modalité flous et la classe inférée.

6. Conclusion :

Dans cet article nous avons présenté une méthode d'extraction de règles de décision depuis des données numériques et nous avons réussi a tiré une connaissance conforme avec l'expert du domaine d'application (cardiologue).

Les arbres de décision flous présentement un avantage majeur dans la classification à cause de leurs simplicité, et leur rapidité d'exécution ainsi de 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 dans le domaine médical, tout expert exige de toute méthode automatique d'aide au diagnostic de justifier ses décisions, une caractéristique absente dans plusieurs techniques citées dans la littérature en particulier les réseaux de neurones. La méthode que nous présentons dans cet article offre aux médecins 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'ajouter d'autres.

La qualité du signal ECG représente une contrainte majeure pour la reconnaissance des différentes pathologies. Ainsi que le mode d'acquisition a un rôle majeur pour différencier entre l'extrasystole ventriculaire et les blocs de conduction. Nos données extraites de la base MIT-BIH est composée essentiellement de battements de la dérivation DII ce qui constitue un handicap majeur lors de la classification.

Nous avons réussi à implémenter un classifieur basé sur l'arbre de décision flou. Les résultats obtenus sont très encourageants, vu le manque d'informations dans la base de données utilisée (présence d'une seule dérivation). Le meilleur classifieur dans les expérimentations menées a un taux de classification de 71%, une performance qui peut être améliorée, si on augmente le nombre de dérivations. Nous avons mené plusieurs changements sur quelques paramètres (choix des nombres des modalités floues et l'emplacement des points modaux) afin de choisir la meilleure structure.

7. References

[1] World Health Organization 2005, Library Cataloguing-in-Publication Data.

[2] Lagerholm, M., and al., "Clustering ECG complexes using hermite functions and self-organizing maps", IEEE Trans. Biomed. Eng. pp. 838-848, 2000.

[3] Silipo1 R., "Investigating electrocardiographic features in fuzzy models for cardiac arrhythmia classification", 4th workshop on intelligent data anlysis in medecine and pharmacology (IDAMAP), Washington, Nov 1999.

[4] Belgacem, N., "détection et classification des arythmies cardiaques par application des réseaux des neurones". juin 2002.

[5] Hedeili, N., "Classification des arythmies cardiaques par l'analyse composante principale et les réseaux de neurones".2004.

[6] Zadeh L.A, "fuzzy sets", Information and Control, 8: 338-353, 1965.

[7] Une interface développée sous matlab, "Laboratoire de Génie biomédical", l'Université de Tlemcen.

[8] Acharya R.U., and al., "Classification of cardiac abnormalities using heart rate signals", Med. Bio. Eng. Comp, Vol. 42 p.288-293, 2004.

[9] Zhou J., "Automatic Determining of Premature Ventricular Contraction Using Quantum Neural Networks", Proc. of the 3rd IEEE Symposium on BioInformatics and bioEngineering BIBE'03, pp. 169-173, 10-12 March 2003.

[10] Serge Guillaume, Brigitte Charnomordic and Jean-Luc Lablée. "FisPro: open source software for systems fuzzy inference"

INRA-Cemagref http://www.inra.fr/bia/M/fispro,2002.

[11] Behadada O., « Application des arbres de décision flous dans la reconnaissance des arythmies cardiaques » 06 décembre 2007.

[12] Serge GUILLAUME, "Représentation des connaissances et systèmes d'inférence floue". THÈSE Doctorat en Génie informatique, Automatique, Traitement du signal. Université Paul Sabatier, Toulouse III. Soutenu le 21 novembre 2005.