Données multidimensionnelles floues et Graphe de similarité

Fuzzy Multidimensional Data and Similarity Graph

A. Nourizadeh[1,2] F. Blanchard[2] B. Delemer[1] M. Herbin[2]
[1] Service Endocrinologie-Diabétologie-Nutrition, CHU Reims
[2] Laboratoire CReSTIC (EA 3804), Université de Reims Champagne-Ardenne

Résumé :
L'analyse de données nécessite une phase exploratoire pour effectuer des rapprochements par similarité et extraire les informations pertinentes. Dans cette communication, nous proposons une approche par graphes sur un échantillon de données. Les graphes sont construits à partir de données floues. Dans une première étape, nous fuzzifions des données multidimensionnelles qualitatives ou quatitatives. Par une opération d'agrégation, nous définissons ensuite un indice de représentativité dans l'échantillon. Le voisinage de chaque donnée est défini par α-coupes. Enfin nous connectons chaque donnée à son voisin le plus représentatif pour construire un graphe dépendant du seuil α utilisé. Cette approche est illustrée par une étude sur l'insulinothérapie dans le cas du diabète de type 2 des personnes âgées.

Mots-clés :
Données multidimensionnelles, Fuzzification, Agrégation, Graphe, Représentants

Abstract:
Data analysis requires an exploratory phase to make connections between similar data and extract relevant information. In this paper, we propose an approach based on graphs of data. The graphs are constructed from fuzzy data. In a first step, we fuzzify both qualitative and quatitative multidimensional data. Then we define an index of representativeness within the data samples using an aggregation operator. The neighborhood of each data is defined by α-cuts. Finally we connect each data to its most representative neighbor to build a graph which depends on the used threshold α. This approach is exemplified to study insulino-therapy in the case of type 2 diabetes in the elderly.

Keywords:
Multidimensional Data, Fuzzification, Agregation, Graph, Representatives

1 Introduction

L'analyse de données nécessite généralement une phase exploratoire pour déterminer le traitement le plus adapté. Dans ce contexte, la logique floue est souvent mise à contribution pour gérer l'incertitude et l'imprécision, elle permet aussi une évaluation plus flexible des similarités entre objets [2]. Classiquement la relation de similarité au sens de Zadeh [1] est symétrique. Dans cette communication, la contrainte de symétrie sera levée : à l'aide d'opérateurs d'agrégation, nous proposons une méthode pratique utilisant le flou pour définir un indice de similarité entre données multidimensionnelles, sans se préoccuper de la propriété de symétrie. Une proposition d'analyse exploratoire des données est ensuite développée en utilisant cet indice de similarité.

À l'instar des méthodes de raisonnement à base de cas [3], l'indice de similarité permet d'extraire de l'échantillon des observations ou données particulières. Ces observations représentent :

– soit des regroupements de données,
– soit des données individuelles (i.e. des cas rares).

Cette communication prolonge nos travaux sur la représentativité des données [8], communication dans laquelle ces observations particulières sont appelées représentants de l'échantillon.

Plusieurs remarques préliminaires peuvent être soulignées dans ce type d'approche de l'analyse exploratoire des données. Les cas rares peuvent être très nombreux dépassant largement 10% de l'effectif total (par exemple 18% dans [11]). Les connaissances a priori sont minimales comme dans le démarrage à froid d'un raisonnement à base de cas ou d'un système de recommandation. Aucun clustering préalable n'est imposé, contrairement aux approches par typicalité [7].

Si l'approche proposée vise à imposer peu de contraintes, en revanche elle nécessite une étape d'interprétation des résultats qui sera propre à chaque application. Comme pour toute analyse exploratoire, une étape de visualisation est

impérative. Cette étape peut être sophistiquée comme dans [6] ou au contraire très simple comme nous le proposons dans cette communication. Nous développons une méthode de construction de graphes de similarité pour structurer l'échantillon et permettre d'observer des connexions induites entre données par notre indice de similarité. Notre objectif est de rendre plus facile l'exploration d'un ensemble des données et de faciliter l'interprétation par des représentations graphiques simplifiées.

2 Données multidimensionnelles et domaine de définition

Soit E un échantillon de n données défini par : $E = \{X_i \ / \ 1 \leq i \leq n\}$. Les données appartiennent à un espace de dimension p. Autrement dit, chaque donnée X appartenant à E a p composantes. Ainsi la donnée X est définie par : $X = (x_r)_{1 \leq r \leq p}$. Les composantes d'une donnée X sont :

– soit quantitatives et définies dans un intervalle de $\mathbb{R}$. La valeur quantitative x_r appartient alors à un domaine D_r où : $D_r = [a_r, b_r]$.
– soit qualitatives. La composante x_r appartient alors à un domaine D_r avec $D_r = \{1, 2, 3, ..., v\}$ où v est le nombre de valeurs que peut prendre x_r.

Le domaine de définition de l'échantillon E est alors défini par : $\Omega = \prod_{1 \leq r \leq p} (D_r)$ où $\prod$ est le produit cartésien des p domaines des composantes.

3 Fuzzification des données

Une valeur qu'elle soit quantitative ou qualitative est souvent imprécise et incertaine et il est classique de la représenter soit par un nombre flou, soit par une quantité floue.
Soit une donnée X de l'échantillon E. Chacune des p composantes x_r de X peut ainsi être représentée par un sous-ensemble flou de son domaine D_r. La fonction d'appartenance à ce sous-ensemble flou est alors telle que :

$$\begin{array}{rccc} \mu_{x_r} : & D_r & \longrightarrow & [0,1] \\ & t & \longmapsto & \mu_{x_r}(t) \end{array} \quad (1)$$

Dans cet article, ces sous-ensembles flous sont normés avec : $\mu_{x_r}(x_r) = 1$.
En utilisant une classique méthode d'agrégation (voir par exemple [4] ou [5]), nous proposons de définir X comme une donnée floue de E dont la fonction d'appartenance μ_X est définie par :

$$\begin{array}{rccc} \mu_X : & E & \longrightarrow & [0,1] \\ & Y & \longmapsto & \mu_X(Y) \end{array} \quad (2)$$

X et Y étant deux observations de E. Si $X = (x_r)$ et $Y = (y_r)$ avec $1 \leq r \leq p$, nous proposons de définir μ_X sur E par :

$$\mu_X(Y) = agreg(\mu_{x_r}(y_r)) \quad (3)$$

avec $agreg$ comme opérateur d'agrégation des p degrés d'appartenance. Pour illustrer cette communication, nous utilisons une approche très empirique où $agreg$ est simplement la moyenne arithmétique.

4 Indice de similarité

Cette fuzzification de la donnée X donne lieu à plusieurs remarques. L'observation X est considérée comme une donnée floue sur E et non sur le domaine Ω. Ce sous-ensemble flou de E est normé car $\mu_X(X) = 1$ pour la méthode $agreg$. Soient deux observations X et Y de E, le sous-ensemble flou associé à X définit une relation valuée de comparaison de Y avec X et :

– Y est similaire à X si $\mu_X(Y) = 1$,
– et Y est dissimilaire à X si $\mu_X(Y) = 0$.

Ainsi μ_X définit un *indice de similarité* à X. La relation de similarité induite sur E n'est pas nécessairement symétrique car $\mu_X(Y)$ n'est pas toujours égal à $\mu_Y(X)$.

5 Indice de représentativité dans l'échantillon

En agrégeant les données floues de E, on définit un sous-ensemble flou dont la fonction d'appar-

tenance est :

$$\mu : \begin{array}{ccc} E & \longrightarrow & [0,1] \\ X & \longmapsto & \mu(X) \end{array} \quad (4)$$

avec :

$$\mu(X) = agreg(\mu_{X_i}(X)) \quad (5)$$

pour $1 \leq i \leq n$ avec la méthode d'agrégation notée $agreg$. Le couple (E, μ) définit alors un échantillon flou de données (sous-ensemble flou de E). Dans cette communication, nous utilisons de nouveau la moyenne arithmétique comme opérateur d'agrégation.
Deux remarques découlent des propriétés élémentaires des opérateurs d'agrégation. Plus l'observation X est similaire aux autres observations de E, plus $\mu(X)$ est proche de 1. Si X était similaire à toutes les données de E, alors on aurait $\mu(X) = 1$. Plus l'observation X est dissimilaire des autres observations de E, plus $\mu(X)$ est proche de 0. Si X était dissimilaire à toutes les données de E, alors on aurait $\mu(X) = 0$. La valeur $\mu(X)$ devient alors un indicateur de similarité de X avec l'ensemble E dans sa globalité. Dans cet article $\mu(X)$ est appelé *un indice de représentativité* de X dans E.

6 Voisinage d'une donnée et graphe de voisinage

Soit X une observation de E considérée comme une donnée floue (voir partie 3) de fonction d'appartenance μ_X. Une α-coupe définit un voisinage de X dans E par :

$$V_\alpha(X) = \{Y \in E / \mu_X(Y) \geq \alpha\} \quad (6)$$

$V_\alpha(X)$ n'est pas vide et contient au moins X. Pour chaque valeur de α, on définit alors un graphe sur E en connectant chaque donnée X à la donnée voisine Z_X ayant la plus grande représentativité :

$$Z_X = \underset{Y \in V_\alpha(X)}{\arg\max}(\mu(Y)) \quad (7)$$

Ce graphe a plusieurs composantes connexes. Nous noterons par k le nombre de composantes connexes. On remarque que, dans chaque composante connexe, il existe une et une seule donnée qui est connectée à elle-même. Ces données connectées à elles-mêmes sont appelées représentants de E.
Si $\alpha = 1$, alors $k = n$, il y a n représentants dans E. Si $\alpha = 0$, alors $k = 1$, il y a un seul représentant dans E (aux cas d'égalités près). Dans cette approche exploratoire de E, α permet de définir les représentants de E.

7 Application

Nous avons appliqué cette approche à des données médicales extraites d'une étude en cours au CHU de Reims [12] et portant sur le diabète de type 2 chez des sujets âgés sous traitement insulinique.

En 2030, l'OMS estime qu'il y aura 438 millions de personnes touchées par le diabète. En France en 2009, plus de 3,5 millions de personnes sont diabétiques dont 26% ont plus de 76 ans [9]. La plupart des systèmes d'aide à l'insulinothérapie concernent le diabète de type 1 [10]. Pour concevoir un système d'aide aux patients diabétiques de type 2, nous avons entrepris une première étude prospective dans le but de modéliser leur insulinothérapie.

Dans cette communication, nous présentons les résultats obtenus sur un échantillon de 44 données (44 sujets diabétiques de type 2, âgés de plus de 65 ans). Les variables sont :
– l'âge en années modélisé par un nombre flou triangulaire ($age - 10$, age, $age + 10$),
– le poids en kg modélisé à l'aide d'une fonction trapèzoïdale ($poids - 10$, $poids - 2$, $poids + 2$, $poids + 10$),
– l'objectif glycémique modélisé par un trapèze dépendant des limites minimales et maximales fixées par l'équipe médicale ($obj_{min} - 0.1$, $obj_{min} + 0.1$, $obj_{max} - 0.1$, $obj_{max} + 0.1$),
– le sexe en variable binaire avec deux valeurs 0 ou 1,
– la dose basale en unités d'insuline modélisée par un trapèze ($basal-5$, $basal-1$, $basal+1$,

$basal + 5$),
- la dose bolus en unités d'insuline modélisée par un trapèze ($bolus-5$, $bolus-1$, $bolus+1$, $bolus + 5$),
- la durée du diabète en années modélisée par un trapèze ($duree - 6$, $duree - 2$, $duree + 2$, $duree + 6$),
- la glycémie HbA1c en % modélisée par un trapèze ($HbA1c-0.5$, $HbA1c-0.2$, $HbA1c+0.2$, $HbA1c + 0.5$),
- la fonction rénale rénale $MDRD$ modélisée par un trapèze ($MDRD - 7$, $MDRD - 2$, $MDRD + 2$, $MDRD + 7$),
- la présence de traitement associé $metvic$ en variable binaire,
- la présence de complication $IDMAVC$ en variable binaire.

La méthode de traitement décrite dans cette communication a été appliquée à cet échantillon. Après l'étape de fuzzification de chaque donnée, nous procédons au calcul de l'indice de représentativité de chaque individu dans l'échantillon, puis à la construction du graphe de similarité pour différentes valeurs du paramètre α. Les figures 1(a), 1(b) et 1(c) représentent les graphes obtenus avec $\alpha = 0.35$, $\alpha = 0.42$ et $\alpha = 0.60$.

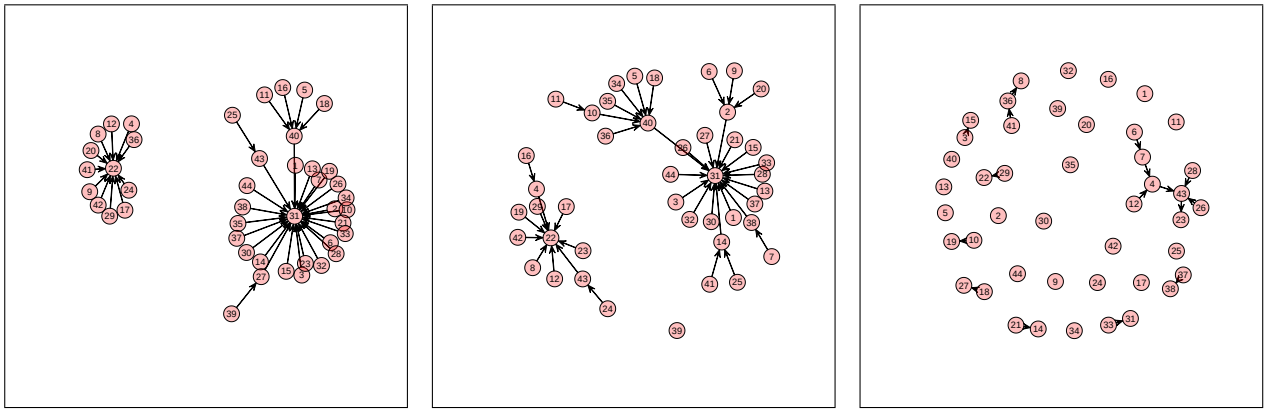
Lorsque α augmente, la taille des voisinages diminue et le nombre de représentants (i.e. le nombre de composantes connexes du graphe) augmente. Le graphique de la figure 2 représente l'évolution du nombre de composantes connexes en fonction des valeurs de α. On constate la propriété triviale suivante :
- lorsque $\alpha = 1$, chaque individu est un représentant et il y a n représentants,
- lorsque $\alpha = 0$, il a un seul représentant dans tout l'ensemble.

Dans cette application, nous ne disposons pas d'information *a priori* nous permettant de choisir une valeur de α. Empiriquement, nous avons fixé $\alpha = 0.42$. Au delà de cette valeur, une faible augmentation de α provoque un morcellement important du graphe. Elle correspond à une sorte de valeur limite au delà de laquelle le graphe de similarité est instable ou trop peu structuré.

Le graphe de la figure 1(b) représente donc le graphe de similarité de notre étude permettant de relier les cas médicaux étudiés. La première information apportée est relative aux arcs orientés du fait de l'indice de similarité non nécessairement symétrique. Ces arcs permettent de rattacher chaque patient à un autre, relativement à sa situation « insulinique ». Le seconde information est celle fournie par le regroupement opéré par les composantes connexes. On obtient ainsi des groupes de patients dont chacun possède un représentant. Cette information est comparable à celle obtenu par les méthodes de *clustering* de type *k-medoïds* [12]. L'intérêt de notre approche est de structurer la base de cas des diabétiques et de permettre au praticien d'observer des analogies et des similitudes parmi les patients de l'étude. Nous sommes bien dans une approche à base de cas.

L'observation du graphe de similarité sur la base de patients diabétiques nous montre qu'un patient (ici représenté par le numéro 39) a une position particulière puisqu'il est un représentant isolé (i.e. une composante connexe constituée d'un seul sommet). Ce type d'information permet d'extraire de la base de cas, les patients dont la situation n'est pas assimilable à celle d'autres sujets, ils constituent des cas atypiques (par exemple : caractéristiques physiques différentes, prescription insulinique particulière) qu'il convient de considérer différemment.

L'étude de ces individus isolés peut être complétée en observant l'évolution de leur nombre et de l'état d'isolement lorsque α varie. La courbe de la figure 3 représente l'évolution du nombre de sommets isolés en fonction de la valeur de α. L'allure de cette courbe est proche de celle représentant le nombre de composantes connexes (le nombre de composantes connexes augmente d'autant plus que le nombre de sommets isolés croît). Les figures 4 et 5 représentent les évolutions des demi-degrés intérieurs, des demi-degrés extérieurs et de l'état d'isolement des individus 39 et 43. Le demi-degré extérieur (courbe bleue) ne peut (par construction) pas être supérieur strictement



(a) $\alpha = 0.35$ (b) $\alpha = 0.42$ (c) $\alpha = 0.60$

Figure 1 – Graphes de similarité pour différentes valeurs de α.

à 1 (on ne relie un sommet qu'à un seul autre). Cette courbe représente donc, pour un individu, la valeur de α à partir de laquelle cet individu devient un représentant. Ainsi dès que les demi-degrés extérieurs et intérieurs sont nuls simultanément, le sommet est isolé. La courbe magenta représente cette information. On peut ainsi constater que l'individu 39 est isolé beaucoup plus tôt que l'individu 43 (i.e. pour une valeur de α plus petite). Ce constat traduit une atypicité plus prononcée de l'individu 39 lorsque l'on compare sa situation avec celle de l'individu 43 (on remarque d'ailleurs que le sommet 39 est isolé dès $\alpha \simeq 0.4$ avant de l'être définitivement à $\alpha \simeq 0.55$).

8 Conclusion

On a défini, d'une part, un graphe à partir d'un échantillon de données multidimensionnelles qualitatives et quantitatives et, d'autre part, une méthode pour extraire des représentants de cet échantillon. La méthode met à contribution le concept de flou pour définir un indice de similarité entre données. La méthode permet de :

- structurer l'échantillon E par un graphe,
- sous-échantillonner E par k représentants,
- faciliter la compréhension de E à l'aide de ces k exemples qui ne sont pas des prototypes virtuels.

Ainsi les représentants peuvent permettre de définir une typologie à l'intérieur de

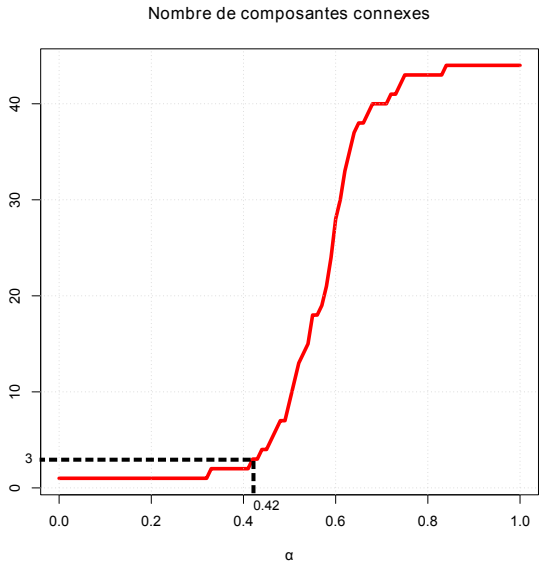


Figure 2 – Nombre de composantes connexes en fonction de α (3 composantes connexes lorsque $\alpha = 0.42$).

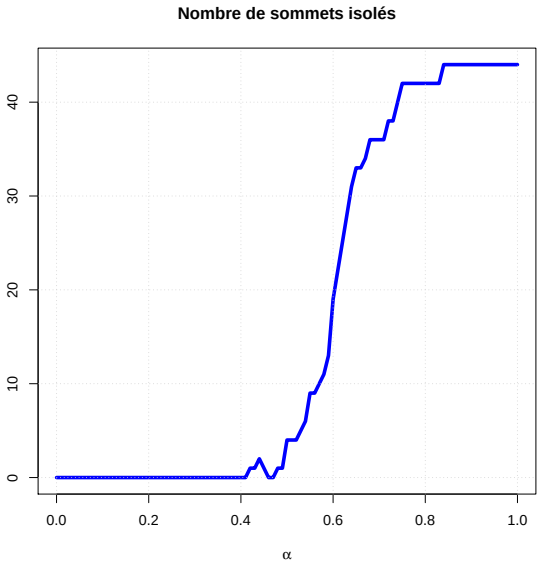


Figure 3 – Nombre de sommets isolés en fonction de α.

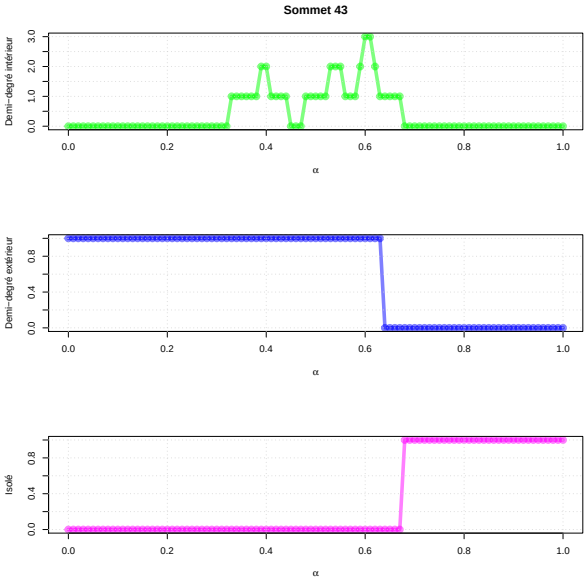


Figure 5 – Demi-degré intérieur, demi-degré extérieur et fonction indicatrice de l'isolement de l'individu 43.

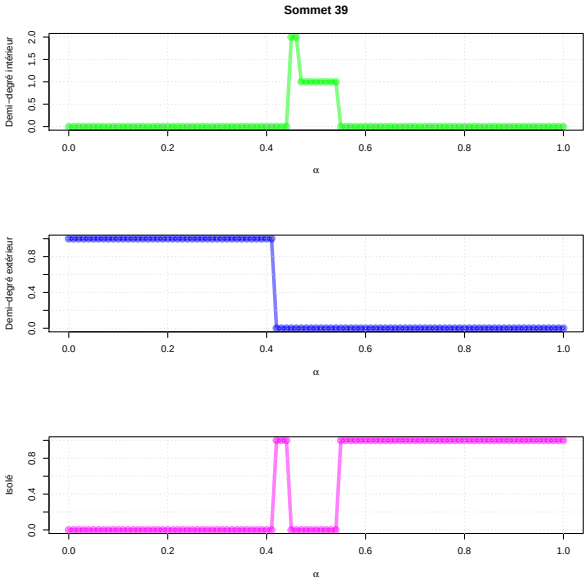


Figure 4 – Demi-degré intérieur, demi-degré extérieur et fonction indicatrice de l'isolement de l'individu 39.

l'échantillon E. Cette approche sans contrainte préalable de clustering ne nécessite pas d'effectif important et s'adapte à l'initialisation d'une base de cas pour un démarrage à froid d'un raisonnement.

Cette approche exploratoire d'un échantillon de données est illustrée par l'utilisation de données médicales en vue de déterminer les différents modèles de traitements insuliniques dans le cas du diabète de type 2 chez des sujets âgés. Ce cas d'étude est complexe car ces patients souffrent souvent de multiples pathologies rendant difficile l'extraction d'une typologie insulinique pour le diabète de type 2. Si l'insulinothérapie est si difficile à modéliser dans le type 2, cela tient à la fois à la variabilité biologique intrinsèque et à la conjugaison des phénomènes d'insulino-résistance et d'insulino-sensibilité spécifique de chaque patient. L'approche que nous proposons est une première étape exploratoire vers la modélisation de l'aide à l'insulinothérapie pour des patients diabétiques âgés.

Remerciements

Les auteurs souhaitent remercier les relecteurs anonymes pour leurs remarques et suggestions pertinentes.

Références

[1] L. Zadeh. Similarity relations and fuzzy orderings. *information Science*, 3 : 177-200, 1971.

[2] W.J. Wang. New similarity measures on fuzzy sets and on elements. *Fuzzy Sets and Systems*, : 305-309, 1997.

[3] P. Bonissone, W. Cheetham. Fuzzy Case-Based Reasoning for Residential Property Valuation. *Handbook on Fuzzy Computing*. Oxford University Press, 1998.

[4] M. Detyniecki. Mathematical aggregation operators and their application to video querying. *Research Report, LIP6, Paris*, 2001.

[5] D. Dubois, H. Prade. On the use of aggregation operations in information fusion processes. *Fuzzy Sets and Systems*, 142 : 143-161, 2004.

[6] F. Blanchard, L. Lucas, M. Herbin. A New Pixel-Oriented Visualization Technique Through Color Image. *Information Visualization*, 4(4), 257-265, 2005.

[7] M.J. Lesot, M. Rifqi, B. Bouchon-Meunier. Fuzzy prototypes : From a cognitive view to a machine learning principle. *Fuzzy Sets and Their Extensions : Representation, Aggregation and Models* Springer : 431-452, 2007.

[8] F. Blanchard, P. Vautrot, H. Akdag, M. Herbin. Data Representativeness Based on Fuzzy Set Theory. *Journal of Uncertain Systems* 4(3), 216-228, 2010.

[9] P. Ricci, P.O. Blotiere, A. Weill, D. Simon, P. Tuppin et al. Diabète traité : quelles évolutions entre 2000 et 2009 en France ? *Bulletin épidémiologique hebdomadaire*, BEH 42-43 : 425-431, novembre 2010.

[10] S. Franc, A. Daoudi, S. Mounier, B. Boucherie, H. Laroye et al. Télémédecine et diabète : état de l'art et perspectives. *Sang Thrombose Vaisseaux*, 23(4) : 178-186, 2011.

[11] M.U. Ahmed, P. Funk. A Computer Aided System for Post-operative Pain Treatment Combining Knowledge Discovery and Case-Based Reasoning. *International Conference Case-Based Reasoning Research and Development*, ICCBR : 3-16, septembre 2012.

[12] A. Nourizadeh, F. Blanchard, A. Aït-Younès, B. Delemer, M. Herbin. Analyse exploratoire de données d'insulinothérapie du diabète de type 2. *JETSAN, Fontainebleau*, 2013.