Possibilistic Pattern Recognition in a Digestive Database for Mining Imperfect Data

ANAS DAHABIAH, JOHN PUENTES, and BASEL SOLAIMAN
TELECOM Bretagne, Département Image et Traitement de l'Information, Brest, France
INSERM, U650, Laboratoire de Traitement de l'Information Médicale, Brest, France
{anas.dahabiah, john.puentes, basel.solaiman}@telecom-bretagne.eu
http://www.telecom-bretagne.eu

Abstract: We propose in this paper a method based, on the one hand, on possibility theory to calculate the similarity among the objects of any casebase, taking into account the imperfection and the heterogeneity of data, and based, on the other hand, on the geometric models like the linear and the circular unidimensional scaling and on the graphic models like the ultrametric trees in order to represent and to visualize this similarity in such a way that we can explore and discover the potential structures and patterns that exist in the data. This approach will be applied to an endoscopic casebase in order to recognize the lesions and the pathologies of this base, and several concrete examples will be given along the paper in order to clarify the mathematical concepts of the method.

Key-Words: - Similarity, Possibility Theory, Linear and Circular Unidimensional Scaling, Ultrametric Trees Endoscopic Images, Data Mining.

1 Introduction

Any mining system whose goal is to analyze or to organize automatically a set of data or knowledge must use, by a way or another, a similarity operator to evaluate the resemblance or the relations that exist in the processed information [1]. For example, measuring the similarity between objects (patient records for example) enables us to gather them into groups in order to understand the characteristics and the behaviour of each group or to classify or to predict the behaviour of a new object. Similarity may also give more efficient organization and retrieval of information, and can simplify our data into more reasonable relationships by using data mining techniques in order to take an action, a plan or a decision. However, the limitations and the restrictions of the traditional similarity measures discussed in section 2 incite us to construct the similarity matrix between the objects of our casebase by means of other tools like the possibility theory measures illustrated in section 3. Similarity estimation based on these measures is explained in details with several numeric examples in section 4. Then, Similarity visualization based on spatial and the graphic representation models is presented in section 5. Our method will be applied on a digestive database consisting of a large number of pathologies (section 6) and we will show that a strong similarity exists between the patient records belonging to the same class of pathology (section 7) and that our method outperforms the prior works (section 8). Then, our results will be discussed and some interesting perspectives will be proposed in section 9.

2 Traditional Similarity Measures Limits

Traditional similarity (dissimilarity) measures (Minkowski, Canberra, Hamming, Jaccard, etc.) [1] suppose generally that the value of each attribute is precise (disregarding the existence of imprecise data), certain (disregarding the existence of uncertain values), and given (disregarding the existence of missing values) while on the contrary, real databases contain a remarkable amount of incomplete, imperfect, and uncertain values. Actually, the uncertainty of data is a delicate widespread problem in many domains. For instance in the medical domain, patients can not describe exactly how they feel or what has happened to them, doctors and nurses can not tell exactly what they observe, laboratories report results only with some degree of errors, physiologists don't precisely understand how the human body works, medical researchers can not precisely characterize how diseases alter the normal functioning of the body, pharmacologists don't fully understand the mechanism accounting for the effectiveness of the drugs, and no one can precisely determine one's prognosis [9]. In addition to their limitations concerning the imperfection of data, the traditional similarity measures have some constraints and conditions that should be considered when dealing with each of them. For instance, division by zero could take place in a considerable amount of these measures, besides we need to know the nature of each variable in the records that contain heterogeneous attributes (quantitative, qualitative, ordinal, etc.) in order to choose

a suitable measure. Moreover, the similarity interval should be taken into account during the aggregation and during the interpretation of the resulting value ([0,1] is the most common similarity interval usually proposed, even though some measures like the angular separation similarity belong to [-1,1]). In reality, a value of an attribute can be given in different ways. For example, if we examine the value of the attribute "age", in some patient records "age" could be assigned as {18 yeas, close to 18 years, more than 15 years, young, between 15 and 20, unknown, 18 or 19, it's quite possible to be 18 or 19 and somehow possible to be 17 or 20, defined by a probability distribution, etc.}. Similarity calculation according to the traditional measures can not be easily carried out between two heterogeneous values, for example, between a value given as 25 and another value given as close to 25, or as a probability distribution, whereas these assignments can be modeled easily thanks to possibility theory [7-8]. For these reasons and in order to construct a general approach, we don't recommend the use of the traditional measures overburdened with a lot of conditions and constraints. Instead, we propose to use the possibility theory measures developed by Zadeh, Prade, Dubois, and Rakoto [2-8] in order to build the similarity (dissimilarity) matrix among the objects of our set.

3 Possibility Theory

Possibility theory provides a method to formalize subjective uncertainties of events, that is to say a means of assessing to what extent the occurrence (the realization) of an event is possible and to what extent we are certain of its occurrence, without having however the possibility to measure the exact probability of this realization because we don't know an analogous event to be referred to, or because the uncertainty is the consequence of observation instrument reliability absence. Let's attribute to each event defined on the universe of discourse Ω (in other words to each element belonging to $\rho(\Omega)$) a coefficient ranging between 0 and 1 assessing to which degree the occurrence of an event is possible, where the value "1" means that the event is completely possible, while the value "0" means that the event is impossible. To define this coefficient, we introduce the possibility measure Π which is a function defined over $\rho(\Omega)$, taking values in $[0,1]$, such that:

Axiom 1: $\Pi(\phi)=0$ (1)

Axiom 2: $\Pi(\Omega)=1$ (2)

Axiom3: $\forall A_1, A_2, \ldots \in \rho(\Omega)$

$$\Pi(\cup_{i=1,2,..} A_i) = SUP_{i=1,2,...}\Pi(A_i) \quad (3)$$

where SUP indicates the supremum of the concerned values.

We can say that the possibility measure is totally defined, if we can attribute a possibility coefficient to all the singletons of Ω. Consequently, the possibility distribution function π defined on Ω, whose values are included in $[0,1]$, such that $\sup_{x\in\chi}\pi(x)=1$ must be defined. As a result the function Π can be defined form the function π by:

$$\forall A \in \rho(\Omega)\ \ \Pi(A)=\sup_{x\in A}\pi(x) \quad (4)$$

Reciprocally, π can be defined form Π by:

$$\forall x \in \Omega\ \ \pi(x)=\Pi(\{x\}) \quad (5)$$

Figure 1 shows some examples of possibility calculation using equation 4:

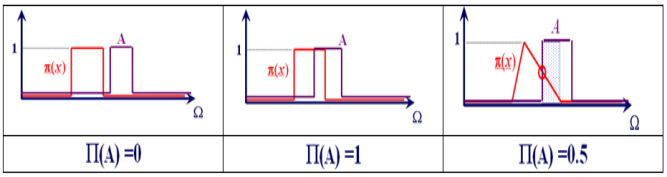

Fig.1 Possibility calculation of an imprecise event *A*.

We should also mention here that the characteristic function of a subset from Ω can be considered as a possibility distribution π defined on Ω. In this case:

$$\Pi(A)=\sup_{x\in\Omega}[\min(\pi(x),\mu_A(x))] \quad (6)$$

Figure 2 Shows an example of this case:

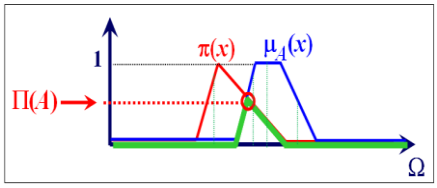

Fig.2 Possibility calculation of a fuzzy event *A*.

To calculate the possibility degree of the couple (x,y) given that $x\in\Omega_1$ and $y\in\Omega_2$ where Ω_1, Ω_2 are two non-interactive universes of discourse, the conjoint possibility distribution defined on the Cartesian product $\Omega_1\times\Omega_2$ should be calculated from:

$$\forall x \in \Omega_1\ \forall y \in \Omega_2\ \ \pi(x,y)=\min(\pi_\chi(x),\pi_\gamma(y)) \quad (7)$$

In fact, the possibility measure is not sufficient to describe the incertitude of the realization of an event, because sometimes the realization of both the event A and its complement A^C could be completely possible simultaneously ($\Pi(A)=1$ and $\Pi(A^C)=1$ at the same time). This means that in this particular case it is impossible to take a decision concerning the realization of A depending on the estimated possibility measure (this case is schematized in figure 3). For this reason, another function, defined on $\rho(\Omega)$, whose values are included in $[0,1]$ and which is called the necessity measure (denoted N) is defined as follows:

Axiom 1: $N(\phi)=0$ (8)

Axiom 2: $N(\Omega)=1$ (9)

Axiom 3: $\forall A_1 \in \rho(\Omega) \quad \forall A_2 \in \rho(\Omega)$

$N(\cap_{i=1,2,....} A_i) = INF_{i=1,2,...} N(A_i)$ (10)

where INF stands for infimum.

There are some interesting relations between the possibility measure Π and the necessity measure N presented in the following equations:

$\forall A \in \rho(\Omega)\ N(A) = 1-\Pi(A^C)$ (11)

$\forall A \in \rho(\Omega)\ N(A) = INF_{x \notin A}(1-\pi(x))$ (12)

$\Pi(A) \geq N(A)$ (13)

$Max(\Pi(a), 1-N(A)) = 1$ (14)

If $N(A) \neq 0$ then $\Pi(A)=1$ (15)

If $\Pi(A) \neq 1$ then $N(A)=0$ (16)

$N(A) \leq \Pr(A) \leq \Pi(A)$ (17)

Where $\Pr(A)$ stands for the probability of any event $A \in \rho(\Omega)$.

Figure 4 gives an example of calculating the necessity degree using equation 11.

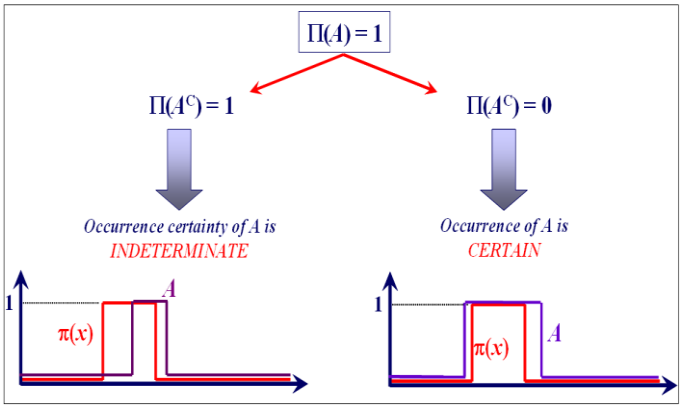

Fig.3 A^C possibility is a good indicator of the occurrence certainty of A.

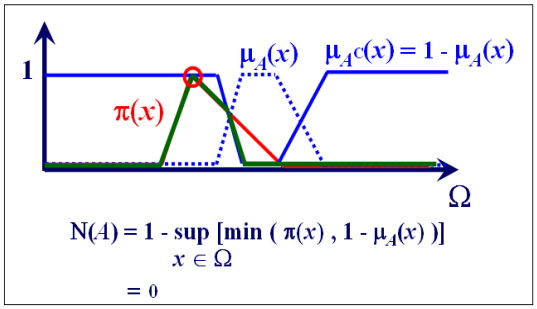

Fig.4 Necessity calculation of a fuzzy event A.

4. Possibilistic Similarity Estimation

Suppose that we have two objects O_j and O_k containing "S" attributes:

$$O_j = [x_{1j} \quad x_{2j} \quad . \quad . \quad x_{ij} \quad . \quad . \quad x_{Sj}]$$
$$O_k = [x_{1k} \quad x_{2k} \quad . \quad . \quad x_{ik} \quad . \quad . \quad x_{Sk}].$$

Each attribute could take a precise or an imprecise value modeled by its possibility distribution, and this value can be either numerical or nominal. The values of some attributes could be unassigned (missing value). Besides, each attribute is associated with a "tolerance function" defined by an expert as a formula or as a table permitting to describe mathematically to which degree we consider that two values of this attribute are similar. An example of tolerance function is the function that we call "close to". Such a function can be defined as:

$$\mu_a(a_x, a_y) = 1 - \frac{|a_x - a_y|}{\Delta} \text{ if } |a_x - a_y| \leq \Delta \quad (18)$$

$\mu_a(a_x, a_y) = 0$ Otherwise

Where Δ is a variable that influences the slope of the function and consequently the notion of "close to". The value of Δ depends on the nature of the attribute and on the user himself. For example, concerning the attribute "age", for an expert the value of Δ might be "2", that means if the difference between two ages is less than "2", they are considered as "similar" with a certain degree of possibility calculated from equation (1), whereas the value of Δ might be "10" for another expert more tolerant. The value of Δ might depends also on the domain of definition of the attribute. For example, for an attribute whose definition domain is $I = [1000000, 2000000]$, the value of Δ might be 1000, whereas for another attribute whose values belong to the interval $I = [0.2, 0.3]$, the value of Δ might be 0.0001. The tolerance function can be also:

- The function of tolerance "True/false": two values of an attribute are similar if they are identical (similarity

equals to 1). If the values are different, the similarity is null, this type of functions is used especially when dealing with nominal variables having independent categories. In the case of ordinal variables we must use the function "close to".

- The "ad hoc" tolerance functions which are defined by the experts to reflect their point of view about the similarities between the attributes.

In our approach the similarity between the two objects O_j and O_k can be estimated by means of two measures: the possibility degree of similarity between O_j and O_k that tells us to which degree it is possible that these vectors are similar, and the necessity degree of similarity of these vectors that tells us to which degree we are certain of their similarity. The probability of the similarity between O_j and O_k exists between the necessity degree that represents the lower limit and the possibility degree that represents the upper limit. To calculate the possibility and the necessity degrees of resemblance, we must calculate the local possibility and necessity degrees between their corresponding attributes and aggregate them by taking their average, for example in order to take a decision concerning the total similarity. The local possibility and necessity degrees of similarity between x_{ij} given by its possibility distribution $\pi_{X_j, x_{ij}}(x_{ij}, y)$ and x_{ik} given by its possibility distribution $\pi_{X_k, x_{ik}}(x, x_{ik})$ for all $i \in \{1,2,...,S\}$ are calculated according to the following relations:

Supposing that D is the definition domain of the considered attribute x_i ($U = D \times D$) and that μ is the tolerance function associated to this attribute, the conjoint possibility distribution π_D is calculated as:

$$\pi_D(x_{ij}, x_{ik}) = \min(\pi_{X_j, x_{ij}}(x), \pi_{X_k, x_{ik}}(y)) \quad (19)$$

In this case, the local possibility degree of similarity π_i can be calculated as:

$$\pi_i(x_{ij}, x_{ik}) = SUP_{u \in U}[\min(\mu(u), \pi_D(u))] \quad (20)$$

The local necessity degree of similarity N_i can be calculated as:

$$N_i(x_{ij}, x_{ik}) = INF_{u \in U}[\max(\mu(u), 1 - \pi_D(u))] \quad (21)$$

We consider that if the value of an attribute is given in one object and is unassigned in the other (the case of missing values), it is completely possible that these values are similar $\pi_i = 1$ but we are entirely uncertain $N_i = 0$ (see figure 5).

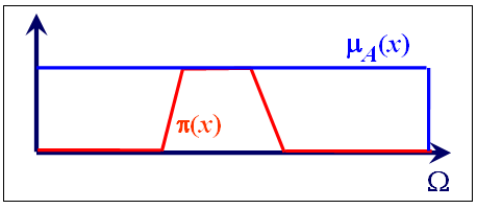

Fig.5 Missing data (total ignorance).

Now that the local possibility and necessity degrees of similarity are calculated, the global possibility and necessity degrees of similarity between X_j and X_k must be determined. In the following we propose different models to aggregate these local degrees:

4-1 Aggregation Using the Average Value

We can take into account all the local degrees of similarity by calculating the average possibility Π_{jk} or to the average necessity N_{jk} as:

$$\Pi_{jk} = \sum_{i=1}^{S} \pi_i \Big/ S \quad (22)$$

$$N_{jk} = \sum_{i=1}^{S} N_i \Big/ S \quad (23)$$

Where *S* is the number of the attributes.

4-2 Aggregation Using the Fuzzy Region Competition

Given that $V = \{v_1, v_2, ..., v_{N_V}\}$ is the set of all the possible values of local possibility (necessity) degrees of similarity and N_V is its cardinality. $V_f = \{f_1, f_2, ..., f_{N_V}\}$ is an ordered set in which each element represents the frequency of the corresponding element of *V* which is calculated as follows:

For *i*=1 to N_V (for each possible value of the local possibility (necessity) degree of the similarity)

$f_i = 0$

For *j*=1 to *S* (for each local possibility (necessity) degree of the similarity)

If π_j is equal to v_i then $f_i = f_i + 1$

Now that we built the set V_f, we create the ordered normalized frequency set $P = \{P_1, P_2, ..., P_{N_V}\}$ of V_f, where P_i is calculated from:

$$P_i = f_i \Big/ \sum_{i=1}^{N_V} f_i \quad (24)$$

Then we apply Dubois-Prade transformation [7-8] in order to construct the possibility distribution set $\Pi = \{\Pi_1, \Pi_2, ..., \Pi_{N_V}\}$ where π_i is calculated from:

$$\pi_i = \max_{l=1,l} \sum_{\{j / \sigma_l^{-1}(j) \le \sigma_l^{-1}(i)\}} p_j \quad \forall i \tag{25}$$

Where σ is a permutation of indexes *{1,2,...,K}* associated to the order $p_{\sigma(1)} \prec p_{\sigma(2)} \prec ... \prec p_{\sigma(K)}$, σ^{-1} is the rank of p_i in the probability list sorted by ascending order, l represents all the possible ascending order sorting when we have at least two equal-value probabilities. The definition domain of the possibility or the necessity degree denoted as $I = [0,1]$ is divided into *C* fuzzy regions whose membership functions are chosen by the user. For example, I might be divided into three regions: the first one represents the most dissimilar records, the second stands for the fairly similar records, and the last one represents the most similar records. Supposing that $\tilde{R}_{\pi k}$ (or $\tilde{R}_{Nk}$ when dealing with the necessity degrees) is the *k-th* fuzzy region and that $\mu_{\tilde{R}_{\pi k}}$ (or $\mu_{\tilde{R}_{Nk}}$ for the necessity) is its membership function ($\mu_{\tilde{R}_{\pi k}} : I \to [0,1]$). We calculate the membership degrees of each element (value) v_i $\forall i \in \{1,2,...,N_V\}$ of the set V to each fuzzy region $\tilde{R}_{\pi j}$ (or to $\tilde{R}_{Nj}$) $\forall j \in \{1,2,...,C\}$, denoted as $\mu_{\tilde{R}_{\pi j}}(v_i)$ (or as $\mu_{\tilde{R}_{Nj}}(v_i)$ for the necessity). For each fuzzy region we calculate the possibility or the necessity membership (the possibility that the similarity between X_j and X_k belongs to the considered region) as follows:

4-2-1 Necessity membership degree (μ_N):

Equation 26 represents the necessity degree that the two records belong to the region R_j given that *V* represents all the possible values of their local possibility degrees, whereas Equation 27 represents the necessity degree that the two records belong to the region R_j given that *V* represents all the possible values of their local necessity degrees:

For *j*=1 to *C* (for all the fuzzy regions)

$$\mu_{N_{JR_{\pi j}}} = INF\left\{ \max_{i=1 to N_V}(\mu_{\tilde{R}_{\pi j}}(v_i), 1 - \Pi_i(v_i)) \right\} \tag{26}$$

$$\mu_{N_{JR_{Nj}}} = INF\left\{ \max_{i=1 to N_V}(\mu_{\tilde{R}_{Nj}}(v_i), 1 - \Pi_i(v_i)) \right\} \tag{27}$$

4-2-2 Possibility membership degree (μ_P):

Equation 28 represents the possibility degree that the two records belong to the region R_j given that *V* represents all the possible values of their local possibility degrees, whereas Equation 29 represents the possibility degree that the two records belong to the region R_j given that *V* represents all the possible values of their local necessity degrees:

For *j*=1 to *C* (for all the fuzzy regions)

$$\mu_{P_{JR_{\pi j}}} = INF\left\{ \max_{i=1 to N_V}(\mu_{\tilde{R}_{\pi j}}(v_i), 1 - \Pi_i(v_i)) \right\} \tag{28}$$

$$\mu_{P_{JR_{Nj}}} = INF\left\{ \max_{i=1 to N_V}(\mu_{\tilde{R}_{Nj}}(v_i), 1 - \Pi_i(v_i)) \right\} \tag{29}$$

We must note that the meaning of the membership degree here is different from that which is used in the fuzzy logic. We consider that the similarity between X_j and X_K belongs to the fuzzy region whose possibility (necessity) membership degree is the maximum. We see here that the decision concerning the similarity can be done in 4 different ways according to $\mu_{N_{JR_{\pi j}}}$, $\mu_{N_{JR_{Nj}}}$, $\mu_{P_{JR_{\pi j}}}$, $\mu_{P_{JR_{Nj}}}$.

4-3 Concrete Examples of Possibilistic Similarity Estimation

Suppose that we would like to calculate the similarity between two patient records in a medical database. Each record contains patient's age, sex, weight, symptoms, biological analysis ...etc. The values of these attributes could be imprecise, vague, uncertain, or unassigned. In all the cases, these values can be easily modeled by possibility distributions. Actually, even if the value of an attribute was assigned as a probability distribution, we are able to transform it to a possibility distribution by means of Prade-Dubois transformation rule [7-8]. For each attribute, we calculate the possibility degree and the necessity degree of similarity between its assigned values in the first and in the second record. We call these degrees "local degrees" since they are estimated at the attribute level. The average degree of all the local degrees calculated between all the considered attributes of the record is called the global degree of similarity

between the records. Let us make things easier by taking numeric values, for this purpose we will take the attribute "age" in the patient record, and will suppose that we consider that the values of two ages are considered similar if the difference between them doesn't exceed ten years old. In other terms, we take the tolerance function (equation 18) whose $\Delta = 10$ (see figure 6). Let us suppose also that the age is assigned in the first record as "is about 40" and in the second record as "is about 50" (see figure 7 in which the value of each age has been modeled by a fuzzy number ± 10).

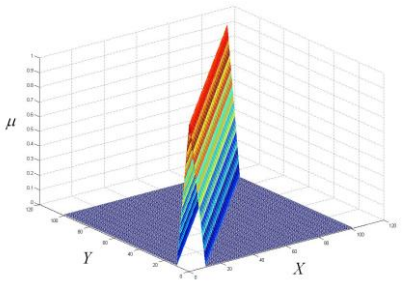

Fig.6 The tolerance function for $\Delta = 10$.

X represents the first fuzzy proposition concerning the value of the attribute in the first object.

Y represents the second fuzzy proposition concerning the value of the attribute in the second object.

μ represents the possibility or the necessity degree.

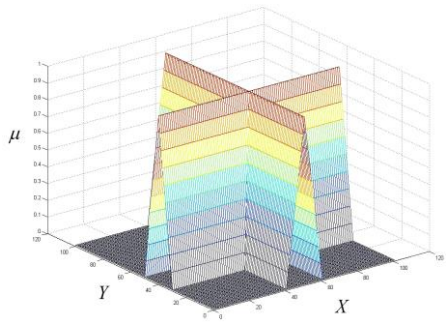

Fig.7 The two ages modeled by fuzzy numbers.

To estimate the local possibility and necessity degrees of similarity, we apply the steps presented above as follows: the conjoint possibility distribution that represents the intersection between the two modeled values of the attribute "age" is calculated using equation 19 (figure 8). The maximum value of the intersection between the tolerance function and the conjoint distribution represents the possibility degree of similarity Π (figures 9 and 10). For the values given in this example we find that $\Pi = 0.60$. Then, we use equation 21 to calculate the necessity degree of similarity (figure 11). We find that $N = 2.22 \times 10^{-16}$. Table 1 shows the local possibility and the necessity degrees of similarity of the attribute "age" for other values of Δ.

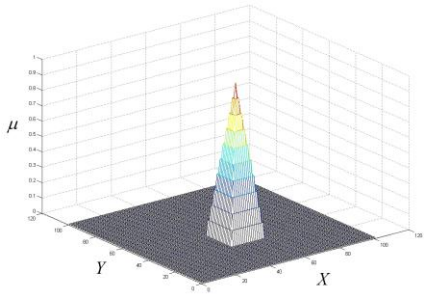

Fig.8 The conjoint possibility distribution.

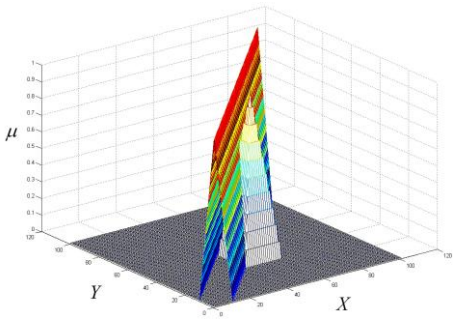

Fig.9 The intersection between the tolerance function and the conjoint possibility distribution.

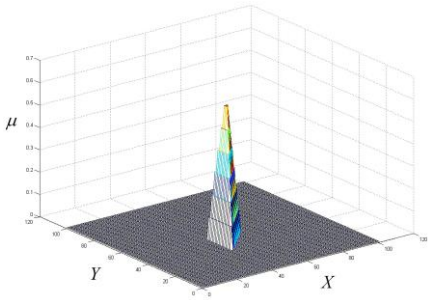

Fig.10 Possibility degree estimation.

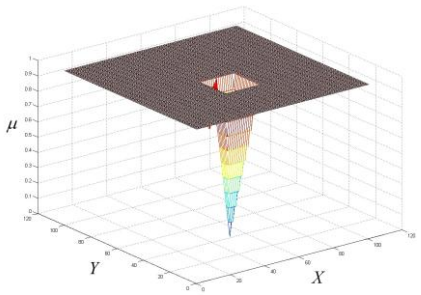

Fig.11 Necessity degree estimation.

Δ	Π	N
20	0.70	0.30
30	0.80	0.40
50	0.84	0.60
99	0.90	0.76

Table 1 Possibility and necessity degrees of similarity of the attribute "age" for different values of Δ

We apply the same steps to all the other attributes of the records taking into account that $\Pi_i = 1$ and $N_i = 0$ if the value of an attribute is assigned in a record and is a missing value in the other record, and that $\Pi_i = 0$ and $N_i = 0$ if the attribute exists in a record and doesn't exist in the other. This can take place in the databases whose records come from different sources (hospitals) because the attributes of the records that come from a hospital can not be exactly the same as those which come from another one even if all the records characterize the same pathology.

Suppose that we have two similar records containing ten attributes, whose local possibility degrees of similarity are {0.90, 0.80, 0.75, 0.70, 0.80, 0.80, 0.75, 0.75, 0.80, 0.70}, and that we have three fuzzy regions {dissimilar, somehow similar, similar}. In order to know to which region belong these two records, we calculate their total necessity membership degrees to each region with the help of table 2 and equation 26 as follows:

$$\mu_{N_{Dissimilar}} = INF\left\{\max_{i=1to10}(\mu_{Dissimilar}(v_i), 1-\Pi_i(v_i))\right\}$$

$$\mu_{N_{Dissimilar}} = Min(0.70,0.70,0,0.90) = 0$$

$$\mu_{N_{SomehowSimilar}} = INF\left\{\max_{i=1to10}(\mu_{SomehowSimilar}(v_i), 1-\Pi_i(v_i))\right\}$$

$$\mu_{N_{SomehowSimilar}} = Min(0.70,0.40,0.20,0.90) = 0.20$$

$$\mu_{N_{Similar}} = INF\left\{\max_{i=1to10}(\mu_{Similar}(v_i), 1-\Pi_i(v_i))\right\}$$

$$\mu_{N_{Similar}} = Min(0.70,0.75,0.80,0.90) = 0.70$$

As we can see, $\mu_{N_{Similar}} \succ \mu_{N_{Somehow}}$ and $\mu_{N_{Similar}} \succ \mu_{N_{Dissimilar}}$. Consequently, we can conclude that the two records are quite similar (as expected).

(In this example: $\Pi_1 = p_1 + p_4$, $\Pi_2 = p_2 + p_1 + p_4$, $\Pi_3 = p_3 + p_2 + p_1 + p_4$, $\Pi_4 = p_4$).

V (all possible π_i)	0.70	0.75	0.80	0.90
p_i	0.20	0.30	0.40	0.10
Π_i	0.30	0.60	1	0.10
$1-\Pi_i$	0.70	0.40	0	0.90
$\mu_{Dissimilar}$	0	0	0	0
$\mu_{SomehowSimilar}$	0.30	0.25	0.20	0.10
$\mu_{Similar}$	0.70	0.75	0.80	0.90
$Max(1-\Pi_i, \mu_{Dissimilar})$	0.70	0.40	0	0.90
$Max(1-\Pi_i, \mu_{SomehowSimilar})$	0.70	0.40	0.20	0.90
$Max(1-\Pi_i, \mu_{Similar})$	0.70	0.75	0.80	0.90

Table 2 Concrete example of estimating the similarity using fuzzy region competition aggregation between two similar records

Suppose now that we have two dissimilar records containing ten attributes, whose local possibility degrees of similarity are {0.10, 0.10, 0.15, 0.30, 0.15, 0.30, 0.15, 0.40, 0.15, 0.15}, and that we have three fuzzy regions {dissimilar, somehow similar, similar}. In order to know to which region belong these records, we calculate their total necessity membership degrees to each region with the help of table 3 and equation 26 as follows:

$$\mu_{N_{Dissimilar}} = INF\left\{\max_{i=1to10}(\mu_{Dissimilar}(v_i), 1-\Pi_i(v_i))\right\}$$

$$\mu_{N_{Dissimilar}} = Min(0.90,0.85,0.70,0.90) = 0.70$$

$$\mu_{N_{SomehowSimilar}} = INF\left\{\max_{i=1to10}(\mu_{SomehowSimilar}(v_i), 1-\Pi_i(v_i))\right\}$$

$$\mu_{N_{SomehowSimilar}} = Min(0.50,0.15,0.50,0.90) = 0.15$$

$$\mu_{N_{Similar}} = Min(0.50,0,0.50,0.90) = 0$$

As we can see, $\mu_{N_{Dissimilar}} \succ \mu_{N_{Somehow}}$ and $\mu_{N_{Dissimilar}} \succ \mu_{N_{Similar}}$. Consequently, we can conclude that the two records are dissimilar (as expected).

(In this example: $\Pi_1 = \max(p_4 + p_1, p_4 + p_3 + p_1)$, $\Pi_2 = p_4 + p_1 + p_3 + p_2$, $\Pi_3 = \max(p_4 + p_1 + p_3, p_4 + p_3)$, $\Pi_4 = p_4$).

V	0.10	0.15	0.30	0.40
p_i	0.20	0.50	0.20	0.10
Π_i	0.50	1	0.50	0.10
$1-\Pi_i$	0.50	0	0.50	0.90
$\mu_{Dissimilar}$	0.90	0.85	0.70	0.60
$\mu_{SomehowSimilar}$	0.10	0.15	0.30	0.40
$\mu_{Similar}$	0	0	0	0
$Max(1-\Pi_i, \mu_{Dissimilar})$	0.90	0.85	0.70	0.90
$Max(1-\Pi_i, \mu_{SomehowSimilar})$	0.50	0.15	0.50	0.90
$Max(1-\Pi_i, \mu_{Similar})$	0.50	0	0.50	0.90

Table 3 Concrete example of estimating the similarity using fuzzy region competition aggregation between two disimilar records

5 Similarity Visualization

Visualization is the process of transforming invisible abstract data, information, and knowledge into a visible display in the form of geometric or graphic representations in order to support tasks such as data analysis, information exploration, trend prediction, pattern detection, rhythm discovery and so [16]. Actually, these representational models give observed events a meaningful interpretation and allow future or unseen events to be anticipated through the process of generalization [16-17]. In order to represent the similarity, we have chosen well-developed mathematical

representational models like the linear and the circular unidimensional scaling (LUS and CUS) [18-21] for similarity spatial (geometric) visualization, then we have chosen the ultrametric trees [22-23] for the graph-based visualization.

5-1 Spatial Visualization

The tasks of linear and circular unidimensional scaling can be defined by the attempt to represent the entries in a symmetric proximity matrix through distances between a set of the locations of the objects defined either along a linear continuum or around a closed, circular continuum. These two scaling tasks are approached through a least-squares optimization strategy based on a combination of combinatorial search and iterative projection techniques [21]. The detailed algorithms that we applied in this paper can be found in the articles [18-19] and [24-25].

5-2 Graphic Visualization

This algorithm [22-23] aims to look for an appropriate tree topology with *m* internal nodes (representing the classes) and *n* terminal nodes (representing the objects) in such a way that the length between two leaves approximates their distance and that the terminal nodes are all equally distant from the root. This type of trees is called ultrametric trees and is widely used in hierarchical clustering. The algorithm that we adapted in this paper (detailed in [23]) consists of two main phases: In the first step a best-fitting ultrametric to our possibilistic matrix is generated by using the standard regression solution to a collection of linear equality and inequality constraints that any ultrametric matrix in a specific equivalence class must satisfy. In the second step, an anti-Robinson matrix is constructed by reordering the rows and the columns of the matrix fitted in the first phase in such a way that the entries within each row and column are non-decreasing moving away from the main diagonal in either direction. This anti-Robinson matrix can easily be represented by an inverted tree called in classification literature "a dendrogram".

6 Tested Base

Our digestive endoscope database [10-15] consists of images, object information, and scene information concerning the upper gastrointestinal tract (esophagus, stomach, and duodenum). The endoscopic findings (pathologies) constitute the objects to be depicted thanks to an exhaustive description mode. Each object is described by 24 attributes with 145 modalities (even 33 attributes with 206 modalities if a sub-object exists), and to each attribute is associated a set of all the possible choices. Owing to the fact that the sub-object features depend on the "non-homogenous state" of the Type feature, there are some other relationships between modalities and feature (for example an object whose Density is "unique" has not a Spatial Organization feature, an object whose Shape is "ring-tube" has not a Minor Axis feature, and so on…) or between modalities of different features (for example, modalities of Relief and Thickness features or modalities of object sizes and sub-object sizes,…). For the scene information, A scene is depicted by a patient profile (the sex and age prevalence features as well as a predefined whole of clinical contexts denoting antecedents, circumstances and symptoms), by the objects (at least one), by eventual spatial relations between objects and the complementary procedures to be envisaged to confirm the disease diagnosis. The attributes of this base could be qualitative, quantitative, or unevaluated (missing values). In our test, we calculated the necessity (the average necessity of similarity of the local necessity degrees of similarity) between a given profile and the others, and then we represented the similarity matrix using the spatial and graphic models of visualization. Our base contains the following pathologies: Dilated lumen, Stenosis, Extrinsic compression, Web, Ring, Hiatal hernia, Food, Blood red (liquid), Blood clot, z-line, spot, Circular Barrett's, Moniliasis, Simple erosion, Ulcer (edge), and Petchial mucosa.

7 Experiments and Results

In order to have a simple and a clear representation of our results, we will show in the following as an example a small subset of cases belonging to our global casebase, keeping in mind that this analysis is applicable to any other case because the approach is general and the matrices that we use are submatrices of the general necessity matrix applied upon all the objects of our global casebase.

Suppose that $CB = \{O_1, O_2, ..., O_{18}\}$ is a casebase consisting of 18 objects (figure 12) where $P_1 = \{O_1, O_2\}$ is the set of the objects whose pathology class is "Dilated Lumen", $P_2 = \{O_3, O_4, O_5, O_6, O_7, O_8, O_9, O_{10}\}$ is the set of the objects whose pathology class is "Stenosis (esophagus)", $P_3 = \{O_{11}, O_{12}, O_{13}, O_{14}\}$ is the set of the objects whose pathology class is "Extrinsic Compression", $P_4 = \{O_{15}\}$ is the set of the objects whose pathology class is "Web-Shape", and $P_5 = \{O_{16}, O_{17}, O_{18}\}$ is the set of objects whose pathology class is "Ring-Shape".

First of all, we construct the possibility-based proximity matrix of the objects of *CB* modeled by the global necessity degree of proximity using equations 20 and 21. Using the algorithm of the LUS explained in details in [18-20] to represent the similarity along a linear continuum, and the algorithm of CUS clearly illustrated

in [20] and [24-25] to represent the similarity in a closed circular continuum, we get the results presented in table 4 for the LUS and in table 5 and figure 13 for the CUS.

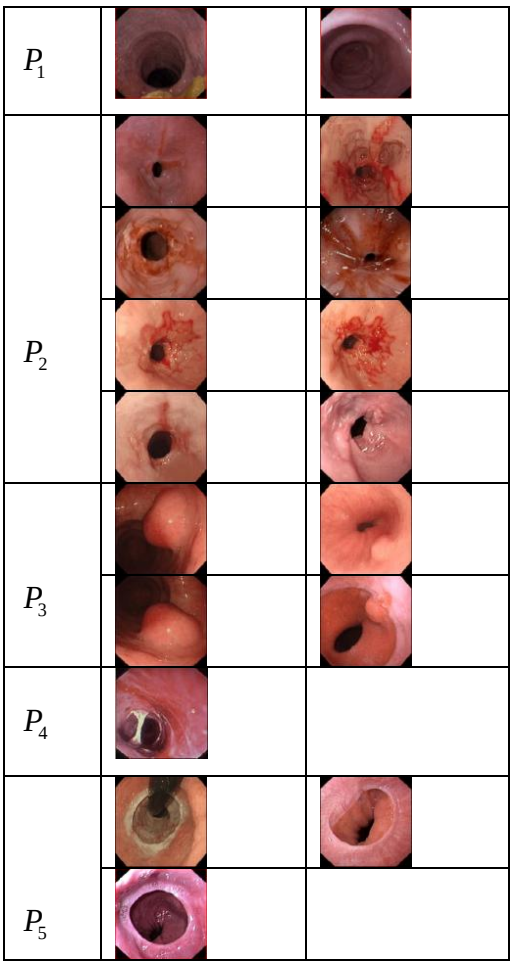

Fig.12 The simplified tested database.

These results show that a strong similarity exists between the objects belonging to the same pathology. In other words, an object belonging to a given pathology is more similar to any other object from the same family than to the other objects belonging to the other pathologies. Thanks to this characteristic, robust retrieval or case diagnostic and reasoning could be achieved here. From the constructed matrix or/and from the obtained categories and coordinates in the table we can study the relationships that exist between the objects belonging to the same class and we can decompose them into other homogeneous groups according to their similarities in order to understand their characteristics or to extract some interesting potential medical rules. Furthermore, we can have an idea about the similarity that exists between the different pathologies. The possibilistic proximity matrix can be also represented as an ultrametric trees using the graph theory techniques (section 5-2), and similar results and conclusion could be obtained (figure 14). Note that the objects belonging to the same pathology are attached to the same internal node. Actually, in our experiments we took all the attributes of each case into consideration with the same importance. However, more interesting and useful results could be obtained by eliminating some useless attributes or by weighting these attributes according to their importance in determining the lesions. In fact, we supposed the unavailability of medical a priori knowledge when we applied our method because we are discussing the general case, nevertheless having some a priori knowledge about the pathologies or about the attributes could lead to more interesting rules and results.

Objects	coordinates	Pathology
O_{15}	-0.7123	P_4
O_{17}	-0.6185	P_5
O_{16}	-0.5382	
O_{18}	-0.4541	
O_1	-0.3733	P_1
O_2	-0.2999	
O_{14}	-0.2012	P_3
O_{12}	-0.1228	
O_{11}	-0.0448	
O_{13}	-0.0004	
O_3	0.1259	P_2
O_{10}	0.2156	
O_7	0.3041	
O_8	0.3758	
O_5	0.4511	
O_9	0.5389	
O_6	0.6255	
O_4	0.7285	

Table 4. *LUS* applied to the dissimilarity matrix of *CB*

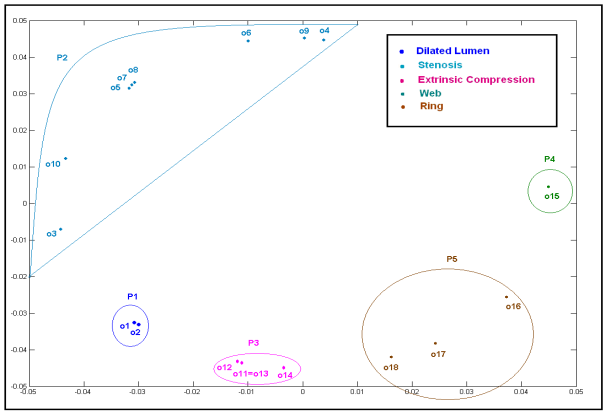

Fig. 13 Circular unidimensional scaling of the possibilistic dissimilarity matrix of *CB*.

Objects	Coordinates X	Y
O_1	-0.0301	-0.0335
O_2	-0.0305	-0.0320
O_3	-0.0444	-0.0073
O_4	0.0037	0.0449
O_5	-0.0313	0.0320
O_6	-0.0101	0.0439
O_7	-0.0309	0.0328
O_8	-0.0301	0.0334
O_9	0	0.0450
O_{10}	-0.0433	0.0124
O_{11}	-0.0111	-0.0436
O_{12}	-0.0121	-0.0432
O_{13}	-0.0111	-0.0436
O_{14}	-0.0037	-0.0449
O_{15}	0.0448	0.0041
O_{16}	0.0371	-0.0256
O_{17}	0.0243	-0.0379
O_{18}	0.0159	-0.0421

Table 5. The coordinates of the points of the *CUS* applied to *CB*.

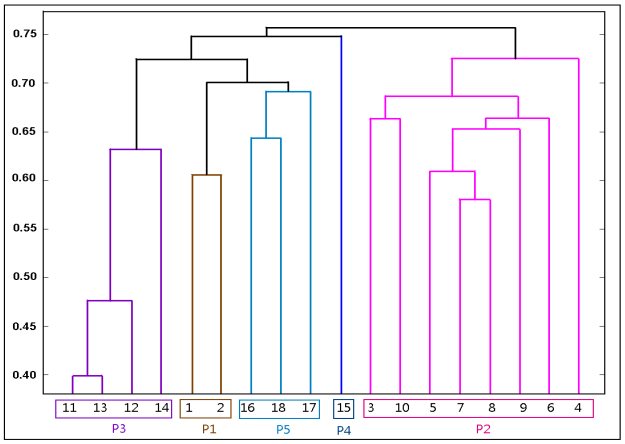

Fig. 14 The ultrametric tree of the possibilistic similarity matrix of *CB* (the horizontal axis represents the index of the objects, while the vertical axis represents the similarity degree).

8 Comparison with Prior Works

Many attempts and methods that aim to overcome the limits and the drawbacks of the traditional measures of similarity have been proposed in the literature. However, these methods have not been general and they treated very particular cases and databases. The most recent and efficient method among them is the method proposed by Zemirline et al. [26-30], presented briefly as follows:

Supposing that Ω is the set of all the modalities of the attributes of the cases in the casebase and that the class (pathology) of each case in this base is known:

For each class and for all the cases belonging to the considered class, the normalized frequency of appearance of each element of Ω is calculated in order to construct this class membership function represented by the histogram. The membership functions of all the classes of the casebase form the knowledge base, from which we calculate the similarity as follows:

Supposing that f_{A_i} is the frequency of appearance of the modality i in the set of cases belonging to the class "*A*", and e_j is the set of the modalities that describe the case j. μ_A is the membership degree to class "*A*" calculated as:

$$\mu_A(f_{A_i}) = f_{A_i} \Big/ \max_{j\in\Omega}(f_{A_j}) \tag{30}$$

The similarity can be calculated by equation 31 (note that the proposed similarity is asymmetric)

$$sim_A(e_i, e_j) = \sum_{k\in e_i, k\in e_j} \mu_A(f_{A_k}) \Big/ \sum_{k\in e_i} \mu_A(f_{A_k}) \tag{31}$$

The major restriction of Zemirline's method is that it supposes that there is a sufficient number of cases that belong to each class in order to build a reliable knowledge base, whereas in reality, sometimes we have only two or three cases of some pathologies in the database, and consequently no reliable membership functions (knowledge base) could be build basing on these objects. Actually, even if we have a considerable number of some cases, nothing can guarantee that these cases represent all the possible models of the considered pathology. Furthermore, this method can not deal with the imperfection of data (imprecision, uncertainty, or the missing values) though this imperfection could change entirely the knowledge base which the authors try to construct. Moreover, this method can not deal with all the types of data that we can find in databases (like the ordinal data for example).

9 Discussion and Perspectives

The proposed approach is very simple and fast since it is based on possibility theory that get use of simple mathematical operations (max, min, addition, …etc). Applying this approach could be very useful in many applications and particularly in data mining since there is no need to hard data pre-processing steps that deal with

the problems of the missing or imprecise values and the heterogeneously-assigned attributes. The simple and the robust way of estimating and visualizing the similarity in all the possible cases could be a strong tool in a large number of data mining techniques like classification, segmentation, clustering, retrieval, etc. and in many other applications that require similarity estimation phase [33]. For instance, this approach enables us to get use of the achieved information stored in the electronic health records whose number is in permanent increase thanks to the cheap storage support and the fast advances in technology [31], in order take the appropriate decision or the correct diagnosis. Furthermore, this method can very used in text-based image retrieval. In fact, image retrieval has been mainly studied based on image content using primitives [32]: color, shape, detected contours, texture, transformation coefficient, etc. In content-based image retrieval there are not imperfect or missing values in the extracted features, while to our knowledge this is the first study that takes into account the imperfect descriptive features of the medical images characterized directly by the doctors or the experts .This method could easily be personalized by taking the user's viewpoints into account through the tolerance function introduced in section 3. Herein our approach has been applied to a digestive database. In the general case, this method could be applied without any modification to any other medical or non-medical database and valuable potential knowledge about the objects could be discovered. As future work, we will study the relations between the attributes of the objects belonging to the same category by studying and comparing their local possibility and necessity degrees in order to extract the key attributes of each pathology and to discover interesting medical rules in the domain of digestive lesions. Besides, we intend to build computer-based training and diagnosis system in this domain.

References:

[1] Bisson, G., *La similarité: une notion symbolique/ numérique*. Apprentissage symbolique-numérique, tome 2. Eds Moulet, Brito. Edition CEPADUES. 2000.

[2] B. Bouchon-Meunier, *La logique floue et ses applications*, Addison Wesley France, chapter 2, 1995.

[3] Dubois, D., Fargier, H., Prade, H., *Possibility theory in constraint satisfaction problems: handling priority, preference and uncertainty*, Applied Intelligence, vol. 6, pp. 287-309, 1996.

[4] Dubois, D., Esteva, F., Garcia, P., Godo, L., Lopez de Mantaras, R., Prade, H., *Fuzzy set-based models in case-based reasoning*, 2nd International Conference on Case-Based Reasoning, ICCBR'97, USA, 1997.

[5] Rakoto, H.; Hermosillo, J.; Ruet, M.; *Integration of experience based decision support in industrial processes*. IEEE conference Page(s):6 pp. vol.7.2002

[6] S. Rashwan, *The Possibilistic Correlation-Dependent Fusion Methods for Optical Detection*, WSEAS transactions on biology and biomedicine, Vol. 4, Issue 5, May 2007.

[7] D. Dubois, L. Foulloy, G. Mauris, and H. Prade. *Probability-possibility transformations, triangular fuzzy sets and probabilistic inequalities*. Reliable Computing, vol. 10 : pp. 273–297, 2004.

[8] D. Dubois, H. Prade, and S. Sandri. On possibility/probability transformations. Proceedings of the Fourth Int. Fuzzy Systems Association World Congress (IFSA'91), Brussels, Belgium, pp. 50–53, 1991.

[9] P. Szolovits. *Uncertainty and decision in medical informatics*, methods of information in medicine, vol. 34, pp. 111-121, 1995.

[10] Le Guillou, C., Cauvin, J-M., *From Endoscopic Imaging and Knowledge to Semantic Formal Images*, Springer, computer science, vol. 4370, pp. 189-201. 2007.

[11] C. Le Guillou, JM. Cauvin, B. Solaiman: *Upper Digestive Endoscopic Scene Analyze*. 23rd Conference of the IEEE Engineering in Medicine and Biology Society, vol. 4, pp. 3855-3858, 2001.

[12] JM. Cauvin, C. Le Guillou, K. Keller: *Similarity and diagnosis in digestive endoscopy*. Proceedings of the 13th World Congress Of Gastroenterology, vol. 7, issue 4, pp. 256-262, 2005.

[13] JM. Cauvin, C. Le Guillou, B. Solaiman: *Computer-assisted diagnosis system in digestive endoscopy*. IEEE Trans Inf Technol Biomed, vol. 7, issue 4, pp. 256-262, 2003.

[14] JM. Cauvin , C. Le Guillou , B. Solaiman: *Diagnostic reasoning by classification in upper digestive tract endoscopy*. World Congress on Medical Physics and Biological Engineering, vol. 1, pp. 31-34, 2000.

[15] http://i3se009d.univ-brest.fr/
Password: view2006

[16] Jin Zhang, *Visualization for Information Retrieval*, Springer Berlin Heidelberg, The Information Retrieval Series Vol.23, pp. 1-20, 2008.

[17] D.J. Navarro, M.D. Lee, *Combining dimensions and features in similarity-based representations Advances in Neural Information*, Processing Systems, vol. 15, pp. 59-66. Cambridge, MA: MIT Press. 2003.

[18] L.J. Hubert, P. Arabie, and J. Meulman, *Linear unidimensional scaling in the L_2-norm: Basic optimization methods using Matlab.* Journal of

Classification, vol. 19, pp. 303–328, 2002.
[19] B. Everitt, D. Howell, *Encyclopedia of statistics in behavioural science*, published by Wiley in 2005.
[20] P. Ian, L. Hubert, J. Rounds; *Circular Unidimensional Scaling: A New Look at Group Differences in Interest Structure*, Journal of counseling psychology, American Psychological Association, vol. 50, n. 3, pp. 297-308, 2003.
[21] M. Brusco, *Optimal least-squares unidimensional scaling: improved branch-and-bound procedures and comparison to dynamic programming*. Springer, vol.70 n2, p253-270, 2005.
[22] A. Dahabiah, J. Puentes, B. Solaiman, Possibilistic ordination-based analysis of an imperfect database, AICCSA proceedings, 2009.
[23] L. Hubert, *Iterative projection strategies for the least-squares fitting of tree structures to proximity data*, British J. Math. Statist. Psych., vol. 48, pp. 281-317, 1995.
[24] L. Hubert, Ph. Arabie, J. Meulman, *Linear and Circular Unidimensional Scaling for Symmetric Proximity Matrices*. British J. Math. Statist. Psych., vol. 50, pp. 253-284, 1997.
[25] L. Hubert, Ph. Arabie, *Iterative Projection Strategies for the Least-Squares Fitting of Tree Structures to Proximity Data*. British journal of mathematical & statistical psychology, British Psychological Society, vol. 48, no2, pp. 281-317, 1995.
[26] A. Zemirline, *Définition et fusion de systèmes diagnostic à l'aide d'un processus de fouille de données : Application aux systèmes diagnostics* (Thesis), Université de Rennes 1, 2008.
[27] L. Lecornu , A. Zemirline, C. Le Guillou,: *Hybrid rule and knowledge based diagnostic system fusion.*. International Journal of Biomedical Engineering and Technology (IJBET), 2007.
[28] A. Zemirline , L. Lecornu , C. Le Guillou :*Data Mining System applied in Endoscopic Image Base*, pp. 1357-1362, ICTTA, 2006.
[29] A. Zemirline , L. Lecornu, B. Solaiman: *Nouvelle méthode d'extraction de règles de classification multi-labels.*. Troisième Atelier Qualité des Données et des Connaissances, Belgique, 2007.
[30] A. Zemirline , L. Lecornu, B. Solaiman. *A new hybrid fusion method for diagnostic systems.* The 11th IPMU International Conference, 2006.
[31] J. Neves, M. Santos, J. Machado, Electronic *Health Records and Decision Support Local and Global Perspectives*, Volume 5, Issue 8, WSEAS transactions on biology and biomedicine, August 2008.
[32] L. Zhang, L. Xi, B. Zhou, *Image Retrieval Method Based on Entropy and Fractal Coding*, WSEAS transactions on systems, vol. 7, Issue 1, 2008.
[33] A. Dahabiah, J. Puentes, B. Solaiman, *Possibilistic evidential clustering*, WSEAS proceedings of AIKED'09, Cambridge University, 2009.