Towards a Possibilistic Processing of Missing Values Under Complex Conditions

Anas DAHABIAH, John PUENTES, and Basel SOLAIMAN
TELECOM Bretagne, Département Image et Traitement de l'Information, Brest, France
{anas.dahabiah, john.puentes, basel.solaiman}@telecom-bretagne.eu
http://www.telecom-bretagne.eu

Abstract: - To estimate the missing values of an attribute in the records of a dataset, all the information provided by the other attributes and the knowledge databases must be considered. However, the information elements could be imperfect (imprecise, possibilistic, probabilistic, etc.) and could have different measuring scales (quantitative, qualitative, ordinal, etc.) at the same time. Furthermore, the relationships and the correlation between the considered attribute and the others should also be pondered. Unlike the prior works that have separately processed these issues using complex and conditional techniques, our approach, essentially based on the tools provided by the possibility theory, can easily handle these aspects within a unified, robust, and simple frameworks. Several numeric examples and applications have been given to simply illustrate the main steps of our method, and some promising perspectives have been proposed at the end of this paper.

Key-Words: - Possibility Theory, Missing Data, Information Imperfection (Uncertainty and Ambiguity), Data Mining.

1 Introduction

The thorny issue of missing values is a problem that continues to plague data mining and knowledge discovery methods and approaches because the majority of mining techniques and algorithms cannot be applied or implemented due to the attributes that include missing data. A common solution of handling missing values is simply to omit from the analysis the attributes or fields with missing contents. Nonetheless, this may be dangerous, since the pattern of missing values may be systematic, and simply deleting objects with missing values would lead to a biased subset of data [1]. Furthermore, it seems like a waste to omit the information in all the other fields, just because one field value is missing [2]. Therefore, data analysts have turned to methods that would replace the missing value with a value substituted according to various criteria. So far, many methods have been developed to deal with the missing data. These approaches have been classified into two main groups: pre-processing methods and the embedded methods [2]. The first ones replace missing values before the data mining process, whereas the second ones deal with them while doing data mining itself. For instance, in [3] the possibilistic similarity that we have proposed can be seen from a certain point of view as an embedded method, because it doesn't require estimating the missing values when measuring the similarity between objects. Instead, it takes account of them during the computation when achieving the other tasks like, clustering, recognition, etc. [3][4]. Nevertheless, in many other applications, the need to estimate the missing values can be indispensible and unavoidable. Accordingly, we will propose in the following another approach that estimates the missing values in the pre-processing phase. Unlike the conventional methods usually dedicated to one type of data measuring scale (qualitative, quantitative, binary, etc.) that neglect the imperfection in the information elements (imprecision, uncertainty, ambiguity, etc.), our approach takes account of all the aforementioned points in a unified framework, by applying a simple, fast, flexible technique, fundamentally based on possibility theory. The next section briefly sums up some previous attempts and works in the domain. Section 3 stresses the deficiency of these works, pointing out the need and the importance of a more sophisticated approach that gets use of the monotone fuzzy measures of possibility theory, briefly presented in section 4. At last, we present our approach that step in to fulfil this need in section 5, followed by two illustrative examples in sections 6 and 7, and some conclusions and remarks in sections 8.

2 Prior Missing Data Methods

Many methods to deal with missing values have been proposed in the literature. These approaches can be classified into two main categories: the pre-processing and the embedded methods [1][2]. Pre-processing methods replace missing values before the data mining process, while embedded approaches deal with missing values while doing data mining itself.

2.1 Pre-Processing Methods

There are two kinds of pre-processing approaches: statistical and machine-learning-based methods. The first kind is simpler and faster and doesn't require complicated processing, while the second type is more accurate and precise but it is time-consuming. In the following we give some examples of the most three applied methods of each type:

2.1.1 Statistical Methods

We present here three well-known approaches. The first one is called *linear regression* [5]. When any two attributes have correlation, we can make an equation of their relationship and predict the missing values by using the equation if either attribute value is known. The second method is called *mean and mode* [6]. The main idea of this approach is to use the mean value for numerical attributes and the most frequent value (the mode) for nominal (qualitative) attributes in the whole dataset to fill up missing values. The last method proposed by [5] aims to fill up missing data by a value that will not disturb each attribute's standard deviation.

2.1.2 Machine-Learning-Based

The most three well-known approaches are: the *nearest neighbor estimator*, the *auto-associative neural networks*, and the *decision tree imputation* [2] [5]. In the first one, we fill up the missing values of an instance by the corresponding values of the nearest neighbor instance. In the second method, an artificial neural network is trained to duplicate all of the inputs as outputs by using back-propagation. When missing values are detected (coded by zeros for example), the network can be used in back-propagation mode. At the input, an appropriate weight can be derived for the missing values so that it doesn't disturb so much the internal structure of the network (or the nonlinear relationship captured by the auto-associative neural network. The last method consists of dividing the main database into two subsets; the first one has no missing data, while the second does. Based on the decision tree extracted from the first subset, we can fill up the attributes with missing data in the second one.

2.2 Embedded Methods

We present here the case-wise deletion method [7], the lazy decision tree approach [8], and C4.5 [9]. The first one ignores the instances that contain unknown attribute values. Lazy decision tree conceptually constructs the best decision tree for each test instance, so if a test instance has a missing value, it makes a decision tree with all attributes in the training dataset except the attribute on which the test instance has missing values. *C4.5* systems induce rule sets in addition to the decision trees. On the training phase, if missing values are occurred at an attribute that is used for branching, *C4.5* creates a new branch called *unknown*. On the testing phase, if a testing instance has missing values; it explores all available branches below the current node and decides the class label by the most probabilistic value. This method assumes that unknown test results are distributed probabilistically in proportion to relative frequent of known result. Figure 1 sums up these methods and techniques:

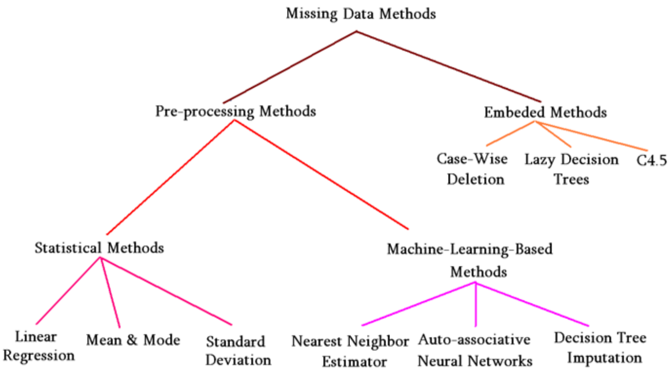


Fig. 1 some prior missing data methods

3 Theoretical Evaluation of the Prior Missing Data Methods

Advances in data acquisition and storage have led to a remarkable growth in datasets and consequently to a high difficulty of the human to analyze or to discover the underlying knowledge or patterns when handling such significant amount of data which can have different measuring scales imperfectly described. For example, the information element "age" in an object like a patient record can take the following values: 25 (precise and quantitative), between 23 and 28 (imprecise and quantitative), young (precise and qualitative), about 25 (imprecise and quantitative), unassigned (missing data), or even given as a probability distribution (probabilistic quantitative or qualitative), etc [3].

Unfortunately, the majority of the till now proposed algorithms deals with crisp values and ignores the existence of imperfect information elements which are frequently confronted in databases. Furthermore, the same dataset could contain different types of attributes (binary, quantitative, qualitative, etc.) at the same time. Nonetheless, these types have been treated separately in the most of cases in the literature [1] [2]. For instance, linear regression, standard deviation, and Nearest neighbor methods can work on quantitative attributes, while decision trees approaches can only deal with the qualitative attributes, and to use this method, quantitative attribute discretization is needed. Along with the aforementioned challenges, there are some technical design problems in some methods that could complicate the estimation. For example, even though the auto-associative neural networks demand a fairly low run time to fill up the missing attributes, the high-dimensionality of the datasets, the choice of the training sets, and the construction of the internal architecture of a network that must be capable to capture the nonlinear relationship is hard and complicated [2]. For these reasons, we will propose in the following a general simple possibility-based missing data estimation approach that can deal with both the crisp and the imperfect values in a heterogeneous database that might contain both numeric and symbolic values under the same framework.

4 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 [4][10].

Let’s attribute to each event defined on the universe of discourse Ω (in other words to each element belonging to the power set $\rho(\Omega)$) 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: $\forall A_1, A_2, ... \in \rho(\Omega)$

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

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:

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

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. 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:

$$\forall x \in \Omega_1\ \forall y \in \Omega_2\ \ \pi(x, y) = \min(\pi_\chi(x), \pi_\gamma(y)) \quad (6)$$

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. 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: $N(\phi)=0$ (7)

Axiom 2: $N(\Omega)=1$ (8)

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

$$N(\cap_{i=1,2,...} A_i) = \inf_{i=1,2,...} N(A_i) \quad (9)$$

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:

$$\forall A \in \rho(\Omega)\ \ N(A) = 1 - \Pi(A^C) \quad (10)$$

$$\forall A \in \rho(\Omega)\ \ N(A) = \inf_{x\notin A}(1-\pi(x)) \quad (11)$$

$$\Pi(A) \geq N(A) \quad (12)$$

$$Max(\Pi(a), 1 - N(A)) = 1 \quad (13)$$

If $N(A) \neq 0$ then $\Pi(A) = 1$ (14)

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

$$N(A) \leq \Pr(A) \leq \Pi(A) \quad (16)$$

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

Possibility theory has lots of very interesting applications in the literature [3-4] [12-13].

4.1 Possibility and Probability Distributions

In some applications, it is sometimes useful to pass from a theoretical platform to another concerning the mathematical models and tools chosen to represent the imperfection in the processed information. To fulfil this need, several useful transformations have been proposed in the literature [11]. In this section, we introduce probability-possibility distribution transformation proposed by Prade and Dubois used in our method [11]. Any probability-possibility transformation must fit the consistency principle informally set by Zadeh as "what is probable is possible", and mathematically interpreted by Dubois and Prade by the inequality: $\Pr(A) \leq \Pi(A)$, $\forall A \subseteq \Omega$, where $\Omega = \{\omega_1, \omega_2, ..., \omega_N\}$, for any possibility or probability measure defined on Ω (in this case we say that Π dominates Pr). Thus, transforming a probability measure into a possibility measure can be materialized by choosing a possibility distribution in $\Im(\Pr)$ (the set of all the possible measures that dominate Pr). Dubois et al have proposed to add the following constraints in order to ensure the preservation of the distribution form: $p_i < p_j \Leftrightarrow \pi_i < \pi_j \quad \forall i, j \in \{1,2,...,N\}$, where $p_i = \Pr(\{\omega_i\})$, and $\pi_i = \Pi(\{\omega_i\})$, for all $i \in \{1,2,...,N\}$. To reduce the imperfection of an information element, the distribution the most specific must be chosen (in the fuzzy set theory we say that the possibility distribution π is more specific than π' if $\pi_i \leq \pi'_i$, $\forall i$). Dubois and Prade show [11] that the solution to this problem exists and it is unique defined as the following:

Supposing that $p_i \neq p_j$, $\forall i$, it is possible to define a strictly ordered relation Ξ on Ω such that: $(\omega_i, \omega_j) \in \Xi \Leftrightarrow p_i < p_j$. Let σ be a permutation of the indices $\{1,2,...,N\}$ associated with the strict order: $p_{\sigma(1)} < p_{\sigma(2)} < ... < p_{\sigma(N)}$, or in another way: $\sigma(i) < \sigma(j) \Leftrightarrow (\omega_{\sigma(i)}, \omega_{\sigma(j)}) \in \Xi$. The permutation σ is a bijection and the inversed transformation σ^{-1} gives the rank of each p_i in the list of the probabilities reordered as an increasing sequence. Accordingly, Dubois-Prade transformation can be given as:

$$\pi_i = \sum_{\{j / \sigma^{-1}(j) \leq \sigma^{-1}(i)\}} p_j \quad \forall i \quad (17)$$

If at least two values of the probability measure are equal, the last equation (proposed for strictly reordered set cannot be applied, because the partially order set P on Ω has to be taken into account. For this purpose, this partial order is represented by a set of its linear extensions Λ(P)={ $\Xi_l, l = 1,2,...,L$ }. At each possible linear extension Ξ_l from Λ(P), there is a permutation σ_l from the set $\{1,2,...,N\}$ that corresponds to Ξ_l in such a way that :

$$\sigma_l(i) < \sigma_l(j) \Leftrightarrow (\omega_{\sigma_l}(i), \omega_{\sigma_l}(j)) \in \Xi_l .$$

In this case, the distribution the most specific and compatible with $\{p_1, p_2, ..., p_N\}$ can be obtained by taking the maximum of all the possible permutations as:

$$\pi_i = \max_{l=1,L} \sum_{\{j / \sigma^{-1}(j) \leq \sigma^{-1}(i)\}} p_j , \ \forall i. \quad (18)$$

For instance, let $p_1 = 0.20, \quad p_2 = 0.50, \quad p_3 = 0.20, \quad and \quad p_4 = 0.10$, there are two possible permutations in this case: $\sigma_1(1) = 4$, $\sigma_1(2) = 1$, $\sigma_1(3) = 3$, and $\sigma_1(4) = 2$; and $\sigma_2(1) = 4$, $\sigma_2(2) = 3$, $\sigma_2(3) = 1$, and $\sigma_2(4) = 2$.

By applying the transformation, we find the following:

$$\pi_1 = \max(p_4 + p_1, p_4 + p_3 + p_1) = \max(0.3, \ 0.5) = 0.5,$$
$$\pi_2 = p_4 + p_1 + p_3 + p_2 = 1,$$
$$\pi_3 = \max(p_4 + p_1 + p_3, p_4 + p_3) = \max(0.5, \ 0.3) = 0.5$$
$$\pi_4 = p_4 = 0.1.$$

Notice that $p_1 = p_3$ implies that $\pi_1 = \pi_3$ (this condition is imposed by the preservation of the strict order).

5 Possibilistic Estimation of Missing Values

Suppose that *DB* is a database that contains N_{DB} objects defined as: $DB = \{D_1, D_2, ..., D_{N_{DB}}\}$. The value of the attribute a_m in a certain object D_m is unknown and has to be estimated based on the given values of this attribute in the other objects that can be heterogeneous, imprecise, probabilistic, etc.

First of all, we gather all the objects which are similar to D_m and in which the value of the attribute a_m is assigned in one set denoted as $DB' = \{D'_1, D'_2, ..., D'_{N_{DB'}}\}$ where $N_{DB'}$ is the number of these similar objects (this can be carried out using the possibilistic similarity measure that we proposed in [3-4] which take in its turn account of information heterogeneity and imperfection (figure 2). a_m can take precise, imprecise, ambiguous, or possibilistic values modeled by a possibility distribution Π_{a_m} in the objects of D'. $V = \{v_1, v_2, ..., v_{N_V}\}$ is the set of all the possible values of a_m in all the records of D' 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 can be obtained as follows:

Suppose that $\mu_{\Pi_{a_m}}(v_i)$ represents the possibility membership degree of the value v_i to the possibility distribution of a_m denoted as Π_{a_m} :

For each possible value of the missing attribute v_i (for *i* = 1 to N_V)

$f_i = 0$, $\mu_{\pi_{v_i}} = 0$

For each object of DB' (for *j* = 1 to $N_{DB'}$)

If $v_i \in \Pi_{a_m}$ with a certain membership degree $\mu_{\Pi_{a_m}}(v_i)$ then $f_i = f_i + \mu_{\Pi_{a_m}}(v_i)$. (19)

Now that we built the set V_f, we create the ordered probability density set $P = \{P_1, P_2, ..., P_{N_V}\}$ where $P_i = f_i \Big/ \sum_{i=1}^{N_V} f_i$ (20). Using the probability-possibility transformation of Prade-Dubois (section 4-1), we can obtain the possibility distribution Π of V_f ($\Pi = \{\Pi_1,\ \Pi_2\ \ ,...,\Pi_{N_V}\}$).

The definition domain of a_m denoted as I_{a_m} (all the possible values of a_m) is divided into *C* fuzzy regions whose widths depend on the nature of this attribute and on the precision required to estimate its value. The membership functions of these fuzzy regions are chosen by an expert who has some a priori knowledge of the attribute a_m.

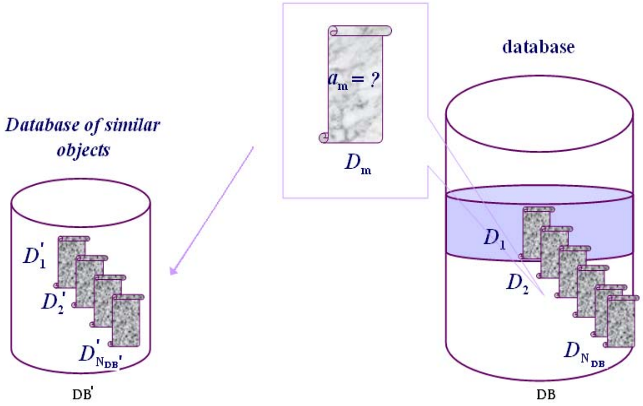


Fig. 2 extracting the similar objects of the dataset

Supposing that $\tilde{R}_{a_m j}$ is the *j-th* fuzzy region of the attribute a_m and that $\mu_{\tilde{R}_{a_m k}}$ is its membership function ($\mu_{\tilde{R}_{a_m k}} : I_{a_m} \to [0,1]$), 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 $\tilde{R}_{a_m j}$, $\forall j \in \{1,2,...., C\}$, denoted as $\mu_{\tilde{R}_{a_m j}}(v_i)$. For each fuzzy region we calculate the possibility or the necessity membership (the possibility that the value of a_m is belonging to the considered region following the next algorithms:

1) Necessity membership degree (μ_N):

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

$$\mu_{N_J} = INF\left\{\underset{i=1toN_V}{\max(\mu_{\tilde{R}_{a_m j}}(v_i), 1-\Pi_i(v_i))}\right\} \quad (21)$$

2) Possibility membership degree(μ_Π):

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

$$\mu_{\Pi_J} = SUP\left\{\underset{i=1toN_V}{\min(\mu_{\tilde{R}_{a_m j}}(v_i), \Pi_i(v_i))}\right\}.$$

We consider that a_m belongs to the fuzzy region whose necessity membership degree is the maximum. If the necessity membership degrees are equal, then we take account of the possibility membership degree.

6 Concrete Example

Suppose that we have an attribute (a_m) with a missing value in a record of a given database taking its value in the interval [0, 10] (the definition domain of a_m), and we want to estimate its value depending on the values of this attribute in the 102 most similar records. Suppose also that in 40 records of these similar records the value is equal to 5 (figure 3-a), in 40 records the value is defined via a possibility distribution depicted in figure 3-b, in other 20 records this value is assigned to be between 7 and 9 (figure 3-c), in a record this value is equal to 7, in the last record this value is equal to 9 (figure 3-d and 3-e). We suppose that the partition of the definition domain of this attribute is designed as depicted in figure 3-f (we have eleven fuzzy regions: around 0, around 1, …, around 10 denoted as *R0, R1, …, R10*), and we must guess to which of these regions the value of the attribute probably belongs.

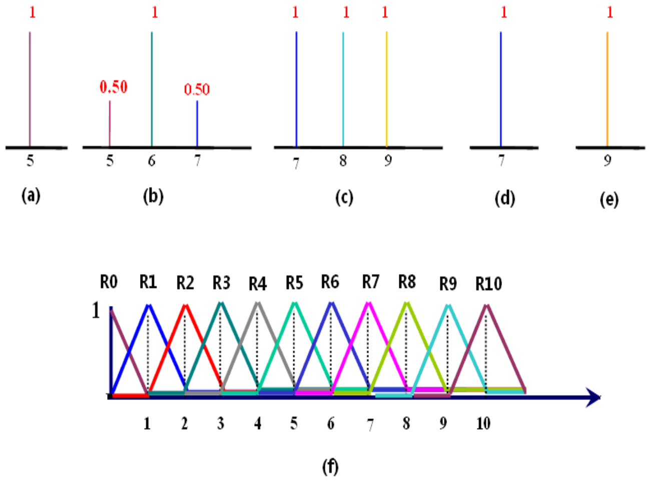


Fig 3. an illustrative example for missing data estimation

According to the steps presented in the last section we can find that:

$V = \{5,\ 6,\ 7,\ 8,\ 9\}$ is the set of all the possible values of this attribute;

$$V_f = \{60,\ 40,\ 41,\ 20,\ 21\},$$

where:

$$v_1 = 1\times 40 + 0.50\times 0.40 + 0\times 20 + 0\times 1 + 0\times 1 = 60$$
$$v_2 = 0\times 40 + 1\times 40 + 0\times 20 + 0\times 1 + 0\times 1 = 40$$
$$v_3 = 0\times 40 + 0.50\times 40 + 1\times 20 + 1\times 1 + 0\times 1 = 41$$
$$v_4 = 0\times 40 + 0\times 40 + 1\times 20 + 0\times 1 + 0\times 1 = 20$$
$$v_5 = 0\times 40 + 0\times 40 + 1\times 20 + 0\times 1 + 1\times 1 = 21$$

Accordingly:

$$P = \{60/182,\ \ 40/182,\ \ 41/182,\ \ 20/182,\ \ 21/182\}$$
$$P = \{0.33,\ \ 0.22,\ \ 0.23,\ \ 0.11,\ \ 0.12\},$$

By applying Prade-Dubois transformation, we find:

$$\Pi = \{1,\ \ 0.45,\ \ 0.68,\ \ 0.11,\ \ 0.23\}$$

According to possibility degrees, we have:

$$\mu_{\Pi_{R5}} = 1,\ \ \mu_{\Pi_{R6}} = 0.45,\ \ \mu_{\Pi_{R7}} = 0.68,\ \ \mu_{\Pi_{R8}} = 0.11,$$
$$\mu_{\Pi_{R9}} = 0.23,\ \ \mu_{\Pi_{Ri}} = 0,\ \ \forall i \in \{0,1,2,3,4,10\}$$

According to necessity degrees, we have:

$$\mu_{N_{R5}} = 0.32,\ \ \mu_{N_{Ri}} = 0,\ \ \forall i \in I/\{5\}.$$

We can sum up that the missing value is about 5, which is intuitive, logic, and expected.

As one might see in this example, the estimation of the missing value is straightforward, simple, and flexible. Thanks to the basic mathematical operations of the processors (minimum, maximum, addition, etc.) on which our approach is based on, the process can be carried out in a fast time. This issue is significantly interesting when handling very large databases as is the case in data mining.

Notice that regardless of the measuring scale of the attributes, all the steps of this technique deal with their possibility degrees. In other words, instead of the numeric value of the attribute supposed in this example (presented on the horizontal axis in the figure), one can take any other type of data, without modifying the calculation.

Remark also that even if the value of the attribute is given via a probabilistic distribution assigned by an automatic system based on other attributes, this distribution can easily be transformed to possibility one using any appropriate transformation in the literature. This can be one of the most complex and challenging points in data missing estimation that cannot be simply handled in the all the previous works and attempts without any constraints, conditions, or prior knowledge.

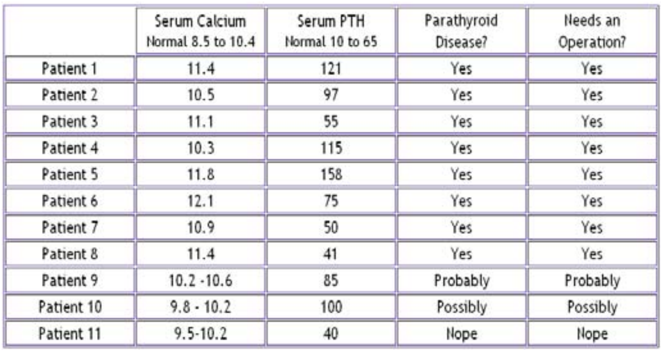
7 Possibilistic Estimation of Correlated Attributes Missing Values

We will show in the following that possibility theory has the potentiality to **easily** solve **complex** situations along with numeric more-developed example. Even if the example deals with simple limited number of attributes and categories, the generality and the robustness of the proposed method are ensured and can easily be proved.

7.1 Problem Description

Let us suppose that among all the features of the records of our dataset, the attributes a_1, a_2, and a_m are connected with some relations between them, and the missing value of a given attribute like a_m (*m* stands for missing) must be estimated by considering the knowledge provided by the attributes a_1 and a_2, given that the attribute definition domains are D_{a_1}, D_{a_2}, and D_{a_m} respectively (the sets of all the possible value of the indexed attributes). For simplicity, it is also assumed that a_m only takes three categorical values C_1, C_2, and C_3, i.e. $D_{a_m} = \{C_1,\ C_2,\ C_3\}$.

This case is very frequent in the everyday real applications. In the medical domain for instance [14], the parathyroid glands situation (a_m) can takes three categorical values ($C_1 = "normal", C_3 = "tumor"$, $C_3 = "Ambiguous"$). Parathyroid tumor can be detected by measuring the hormone PTH (a_1) that the parathyroid glands make and compare this level to the amount of calcium in the blood (a_2). All endocrine glands make hormones, and all hormones have a normal level in our blood. If an endocrine gland develops into a tumor, it will over-produce its hormone. The hormone has effects on other parts of the body. In the case of a parathyroid gland tumor, it overproduces PTH which in turns takes calcium out of the bones and puts it into the blood. It is the high calcium in the blood that makes us sick. i.e. if the blood calcium level is high, it should be associated with a low parathyroid hormone level if the parathyroids are normal. If the blood calcium level is too high, and is associated with a high parathyroid hormone level must be due to a tumor in the parathyroid gland. That is, the high blood calcium is a result of the excess parathyroid hormone (PTH). Table 1 shows examples of patient's calcium levels, PTH levels, and whether or not they have hyperparathyroidism and whether or not they need surgery to remove a parathyroid tumor.



	Serum Calcium Normal 8.5 to 10.4	Serum PTH Normal 10 to 65	Parathyroid Disease?	Needs an Operation?
Patient 1	11.4	121	Yes	Yes
Patient 2	10.5	97	Yes	Yes
Patient 3	11.1	55	Yes	Yes
Patient 4	10.3	115	Yes	Yes
Patient 5	11.8	158	Yes	Yes
Patient 6	12.1	75	Yes	Yes
Patient 7	10.9	50	Yes	Yes
Patient 8	11.4	41	Yes	Yes
Patient 9	10.2 -10.6	85	Probably	Probably
Patient 10	9.8 - 10.2	100	Possibly	Possibly
Patient 11	9.5-10.2	40	Nope	Nope

Table 1. Examples of patient's records having parathyroid tumor [14]

In reality, the influence of the attributes a_1 and a_2 on a_m can be more complicated to describe and estimate. For instance, the influence of the calcium level on the situation of the parathyroid is schematized in figure 4.

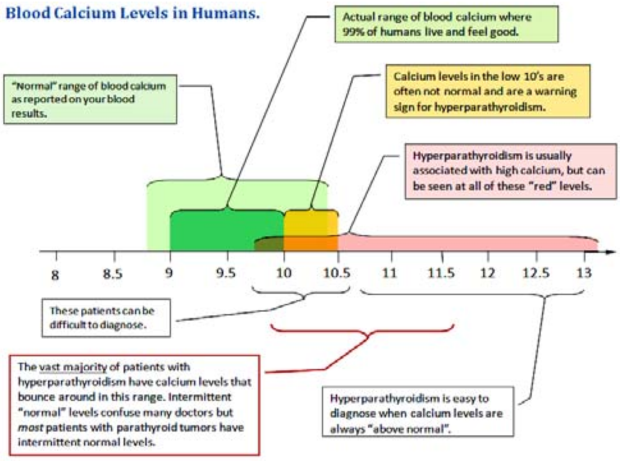

Blood Calcium Levels in Humans.
Actual range of blood calcium where 99% of humans live and feel good.
"Normal" range of blood calcium as reported on your blood results.
Calcium levels in the low 10's are often not normal and are a warning sign for hyperparathyroidism.
Hyperparathyroidism is usually associated with high calcium, but can be seen at all of these "red" levels.
8
8.5
9
9.5
10
10.5
11
11.5
12
12.5
13
These patients can be difficult to diagnose.
The vast majority of patients with hyperparathyroidism have calcium levels that bounce around in this range. Intermittent "normal" levels confuse many doctors but most patients with parathyroid tumors have intermittent normal levels.
Hyperparathyroidism is easy to diagnose when calcium levels are always "above normal".

Fig. 4 the relation between calcium level and parathyroid situation [14]

7.2 Problem Solution

The solution of such problem consists of three straightforward main phases:

- *Knowledge Extraction*: in this step we accumulate all the knowledge provided by the assigned values of the influencing attributes a_1 and a_2, supposing that each of them is a source of information.

- *Possibility Degrees Calculation*: these degrees are computed according to the actual values of a_1 and a_2, via the extracted knowledge in each source.

- The *fusion* of the possibility degrees of the two sources evaluated in the last step.

In the following, these phases are explained in detail:

7.2.1 Step 1: Knowledge Extraction

As both a_1 and a_2 influence the value of a_m that takes three categorical values C_1, C_2, and C_3, we try to know the nature of this influence by depicting the probability distributions of each class according to all the possible values of the attributes a_1 and a_2. This can be known via a knowledge database or it can be computed via the following the algorithm:

```
From all the records of the dataset in which the values of $a_1$, $a_2$, and $a_m$ are given do:
 For each influencing attribute ($a_1$ than $a_2$) do:
  For each category of $a_m$ ($C_1$, $C_2$, and then $C_3$) do:
   For each value of the influencing attribute domain ($D_{a_1}$, and then $D_{a_2}$) do:
Compute "$q_{V_{a_i}/C_j}$" the number of simultaneous occurrences of this value ($V_{a_1}$ or $V_{a_2}$) and the considered category ($C_1$, $C_2$, or $C_3$).
```

The probability that the attribute's value V_{a_i} ($i \in \{1,2\}$ in this example) belongs to the category C_j (where $j \in \{1, \ 2, \ 3\}$ in this example), given that each influencing attribute a_i represents a source of information S_i, can be estimated as:

$$p_{S_i}(V_{a_i}/C_j) = q_{V_{a_i}/C_j} \big/ Q \tag{23}$$

where Q is the total number of the occurrences of all the possible values of D_{a_i}.

Let's explain the precedent algorithm by a simple numeric example:

Assume that the value of the attribute a_m that takes three categorical values C_1, C_2, and C_3 is correlated with the value of an attribute a whose definition domain is defined as: $D_a = \{1, \ 2, \ 3, \ 4\}$. The frequency of occurrence of each value of D_a with the first category C_1 is assumed to be given as $q_{V_a/C_1} = \{10, \ 8, \ 5, \ 2\}$, with C_2 and C_3 it is given as $q_{V_a/C_2} = \{8, \ 10, \ 5, \ 2\}$ and $q_{V_a/C_3} = \{2, \ 5, \ 8, \ 10\}$ respectively. The probability distributions resulted from the information source a can be evaluated using the precedent equation and can be depicted as in figure 5:

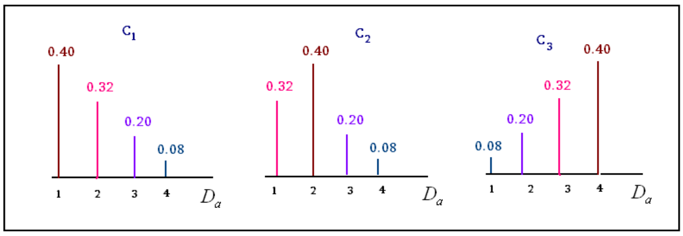


Fig. 5 the probability distributions of the three categories extracted from the influencing attribute

7.2.2 Step 2: Possibility Measures Evaluation

The possibility measures of a given value of the attribute a_i denoted as V_{a_i} to a certain class C_j can be computed from the possibility measures $p_{S_i}(V_{a_i}/C_j)$ calculated in the precedent step as:

$$\Pi_{S_i}(C_j/V_{a_i}) = \frac{P_{S_i}(V_{a_i}/C_j)}{\max\limits_{Category\,k} P_{S_i}(V_{a_i}/C_k)} \tag{24}$$

For instance, in the simple example given above, assume that we want to calculate attribute a_m possibility measure to all its possible categories, given that that $V_a = 1$:

$$\Pi_{S_a}(C_1/V_a=1) = \frac{P_{S_a}(V_a=1/C_1)}{\max\,[P_{S_a}(V_a=1/C_1), \ P_{S_a}(V_a=1/C_2), \ P_{S_a}(V_a=1/C_3)]}$$

$$\Pi_{S_a}(C_1/V_a=1) = \frac{0.40}{\max(0.40, \ \ 0.32, \ \ 0.08)} = 1$$

$$\Pi_{S_a}(C_2/V_a=1) = \frac{P_{S_a}(V_a=1/C_2)}{\max\,[P_{S_a}(V_a=1/C_1), \ P_{S_a}(V_a=1/C_2), \ P_{S_a}(V_a=1/C_3)]}$$

$$\Pi_{S_a}(C_2/V_a=1) = \frac{0.32}{0.40} = 0.80$$

$$\Pi_{S_a}(C_3/V_a=1) = \frac{P_{S_a}(V_a=1/C_3)}{\max[P_{S_a}(V_a=1/C_1), \ P_{S_a}(V_a=1/C_2), \ P_{S_a}(V_a=1/C_3)]}$$

$$\Pi_{S_a}(C_3/V_a=1)=\frac{0.08}{0.40}=0.20$$

Accordingly, we assign the missing value of a_m to the categorical value C_1, because we only have one information source a.

7.2.3 Step 3: The Fusion of the Information Sources

Sometimes, unlike the previous example, we could have several influencing attributes (several information sources). For instance if the missing attribute is the "parathyroid situation" in the medical example previously presented, then we have two influencing attributes in the patient record "the calcium rate" and "the Harmon's rate". In such cases, the possibility measures are calculated form each information source, and then are combined using a suitable fusion operator like the conjunctive fusion which takes the normalized intersection of the possibility distributions resulting from the different sources, assuming that all these sources are consonant (they don't disagree or they disagree slightly) [15] [16]. In this case:

$$\Pi_{S_{12\ldots}}(C_j/(V_{a_1},V_{a_2},\ldots))= \qquad (24)$$

$$\frac{\min[\Pi_{s_1}(C_j/V_{a_1}),\ \Pi_{s_2}(C_j/V_{a_2}),\ldots..]}{\max\left[\min[\Pi_{s_1}(C_1/V_{a_1}),\ \Pi_{s_2}(C_1/V_{a_2}),\ldots..],\ \min[\Pi_{s_1}(C_2/V_{a_1}),\ \Pi_{s_2}(C_2/V_{a_2}),\ldots..],\ldots\right]}$$

Let us suppose for instance that the missing value of the attribute a_m in the aforementioned simple example is correlated with two attributes (the attribute a as in the example, and the attribute b whose definition domain is $D_b=\{1,\ 2\ ,3,\ 4\}$, and which take takes the value $V_b=2$). In this case, we also suppose that the probability distributions extracted from this source are defined as in the figure 6:

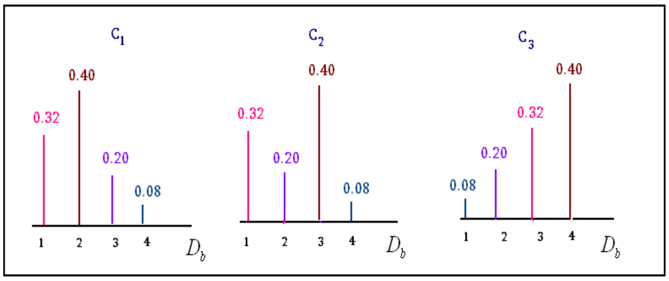


Fig. 6 the probability distributions of the three categories extracted from the second information source

Similarly, the possibility measures resulting from the second information source (attribute b) are computed as:

$$\Pi_{S_b}(C_1/V_b=2)=\frac{P_{S_b}(V_b=2/C_1)}{\max\,[P_{S_b}(V_b=2\,/C_1),\ P_{S_b}(V_b=2\,/C_2),\ P_{S_b}(V_b=2\,/C_3)]}$$

$$\Pi_{S_b}(C_1/V_b=2)=\frac{0.40}{\max(0.40,\ \ 0.20,\ \ 0.20)}=1$$

$$\Pi_{S_b}(C_2/V_b=2)=\frac{P_{S_b}(V_b=2/C_2)}{\max\,[P_{S_b}(V_b=2\,/C_1),\ P_{S_b}(V_b=2\,/C_2),\ P_{S_b}(V_b=2\,/C_3)]}$$

$$\Pi_{S_b}(C_2/V_b=2)=\frac{0.20}{0.40}=0.50$$

$$\Pi_{S_b}(C_3/V_b=2)=\frac{P_{S_b}(V_b=2/C_3)}{\max\,[P_{S_b}(V_b=2\,/C_1),\ P_{S_b}(V_b=2\,/C_2),\ P_{S_b}(V_b=2\,/C_3)]}$$

$$\Pi_{S_b}(C_3/V_b=2)=\frac{0.20}{0.40}=0.50$$

The Fusion of these measures and the measures computed according to the first information source are combined using the disjunctive fusion as depicted in figure 7:

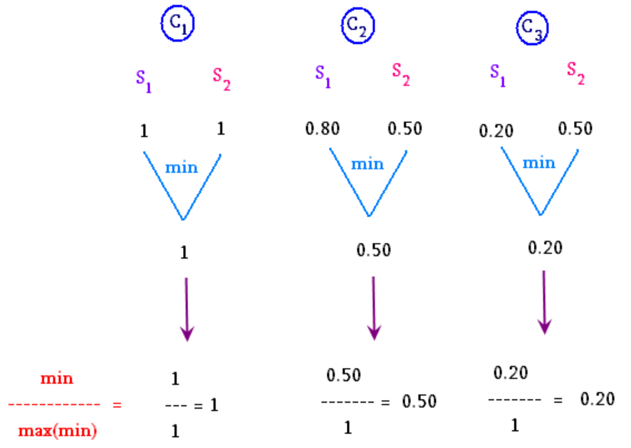


Fig. 7 the fusion of the possibility measures

It can be deduced from the obtained results that according to the given values of the attributes a and b ($V_a=1$, and $V_b=2$), and according to the knowledge extracted from them concerning the relations between them and the attribute a_m whose value is missing (modeled by probability distributions), the most possible value of the missing data is "C_1". Figure 8 resumes and schematizes the main steps to estimate the missing value of an attribute a_m whose definition domain is

$D_m = \{C_1, \ C_2, \ C_3\}$, and which is related to the two attributes a_1 and a_2.

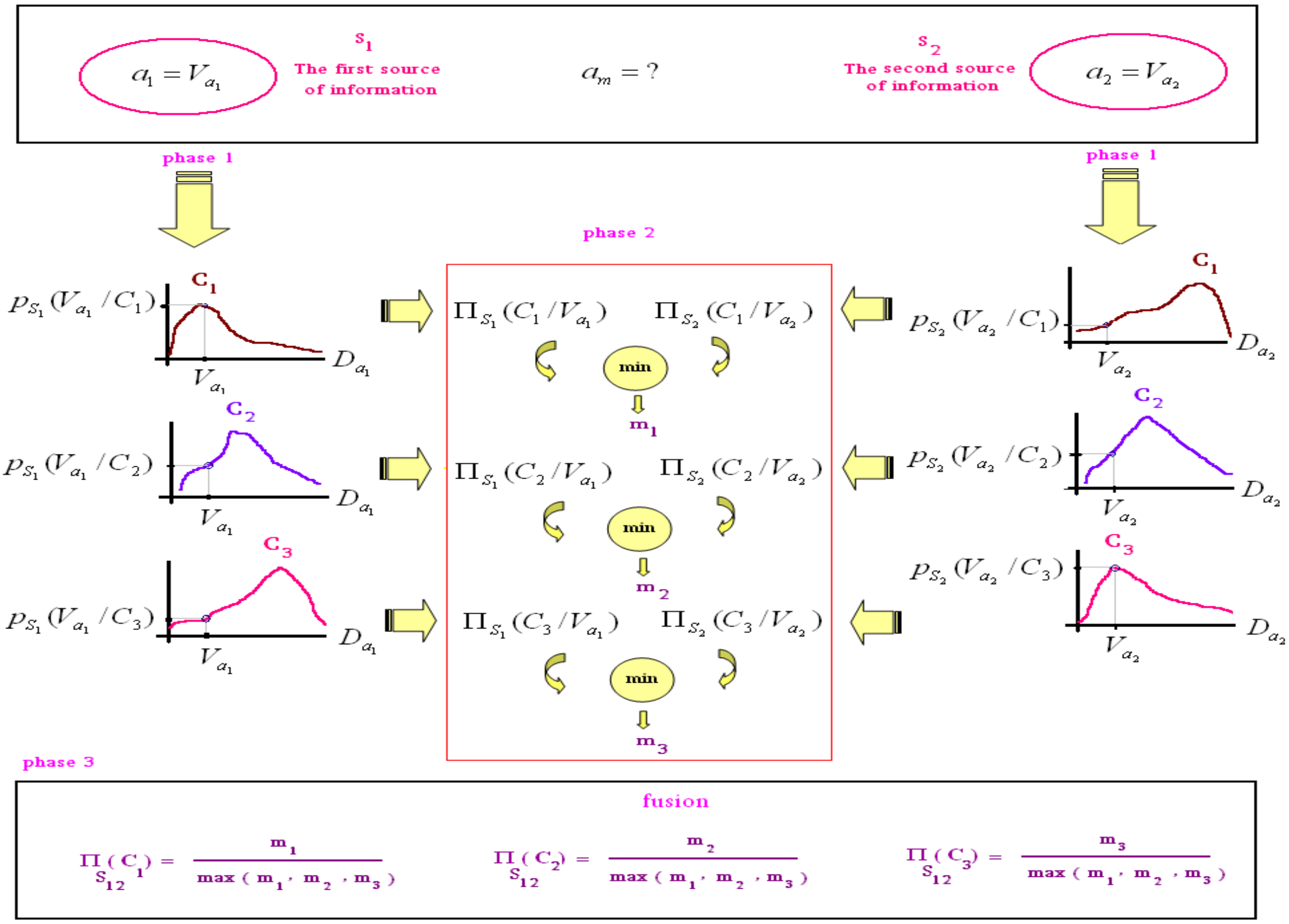


Fig. 8 the main phases of estimating a missing value

8 Conclusion and Perspectives

In this paper, we proposed a possibilistic approach to deal with the heterogeneity and the imperfection of information elements in a unified framework when estimating the missing data. This aspect was neglected, or superficially treated in the literature in spite of its importance in real datasets, and though the estimation of the missing data is essential to accomplish a considerable number of data mining tasks. The proposed approach is simple, straightforward, and can be accomplished in a short time since it is fundamentally based on basic operations (like the maximum, the minimum, the addition, etc.). An illustrative example has been given to simply explain the basic steps of the proposed technique. In spite of its simplicity, this example can be applied to a large spectrum of applications and problems without significant modifications.

Another example that handles the correlation of the attributes in addition to the imperfection and the heterogeneity has been given to show the flexibility of the possibilistic tools in adapting complex conditions and constraints in data mining. More complex situations can be solved in the same way in any other applications.

The potentiality of the proposed approaches to easily deal with all the states of information elements within a unified framework can play a pivotal role in analyzing and mining large real databases, in which the objects consist of a notable number of attributes with considerable variety, and can noticeably reduce the processing time which is an important issue in data mining. In addition, the proposed strategy can be very useful when extracting knowledge databases from imperfectly-described complex objects. Furthermore, it overcomes the obstacles and the challenges encountered in case-based reasoning systems when the cases and their associated solutions are provided via imperfect heterogeneous information elements.

References:

[1] Larose, D., Discovering Knowledge in Data: an Introduction to Data Mining, *Wiley*, pp. 163-179, (2004).

[2] Fujikawa, Y., Ho, T., Cluster-Based Algorithms for Dealing with Missing Values, *LNCS Springer*, vol. 2336, pp. 549-554, (2002).

[3] Dahabiah, A., Puentes, J., Solaiman, B., Possibilistic Pattern Recognition in a Digestive Database for Mining Imperfect Data. *WSEAS Transactions on Systems*, vol. 8, no. 2, pp. 229-240, (2009).

[4] Dahabiah, A., Puentes, J., Solaiman, B., Digestive database evidential clustering based on possibility theory. *WSEAS transactions on biology and biomedicine*, vol. 5, no. 9, pp. 239-248, (2008).

[5] Pyle, D., Data Preparation for Data Mining, *Morgan Kaufmann Publishers*, (1999).

[6] Han, J., Kamber, M., Data Mining: Concepts and Techniques, *Morgan Kaufmann Publishers*, (2001).

[7] Liu, W.Z., White, A.P., Thompson S.G., Techniques for Dealing with Missing Values in Classification, *LNCS Springer*, vol. 1280, pp. 527- 536, (1997)

[8] Friedman, J. H., Khavi, R., Yun, Y., Lazy Decision Trees, *Proceedings of the 13th National Conference on Artificial Intelligence, MIT Press*, pp. 717-724, (1996).

[9] Quinlan, R., C4.5: Programs for Machine Learning, *Morgan Kaufmann Publishers*, (1998).

[10] Bouchon-Meunier, B., Uncertainty Management in Medical Applications, Chapter 1 Nonlinear Biomedical Signal Processing, *Akay, M., (ed.)*, IEEE Press, (2000).

[11] Masson, M.H., Denoeux, T., Inferring a Possibility Distribution from Empirical Data, *Fuzzy Sets and Systems*, vol. 157, no. 3, pp. 319-340, (2006).

[12] A. Dahabiah, J. Puentes, B. Solaiman, *Digestive Casebase Mining Based on Possibility Theory and Linear Unidimensional Scaling*. 8th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, pp. 218-223, (2009).

[13] A. Dahabiah, J. Puentes, B. Solaiman, *Possibilistic Evidential Clustering*. 8th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, pp. 212-217, (2009).

[14] http://www.parathyroid.com/

[15] J. Desachy, L. Roux, E. Zahzah, *Numeric and Symbolic Data Fusion: A Soft Computing Approach to Remote Sensing Images Analysis*, Pattern Recognition Letters, vol. 17, no.13, pp. 1361-1378, (1996).

[16] S. DeveughMe, B. Dubuisson, *Estimation of Geometric Tokens: Possibility Theory Handles Severe Conflicts*, Proceedings of the 8th Scandinavian Conference on Image Analysis, vol. 2, pp. 1365-1372, (1993).