|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FG-AI4H-E-022 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | Plenary | Geneva, 30 May – 1 June 2019 | |
| **DOCUMENT** | | | | |
| **Source:** | | Ministry of Telecommunication and IT Damascus (Syria) | | |
| **Title:** | | Proposing a unified mathematical framework and data mining algorithms to conduct all types of imperfection in medical data | | |
| **Purpose:** | | Discussion | | |
| **Contact:** | | Anas Dahabiah Director of planning and international cooperation in the ministry of telecommunication and IT Damascus; Syria | | Tel: +963931823940 Fax: +963112214578 Email: [adahabiah@moct.gov.sy](mailto:adahabiah@moct.gov.sy) |

|  |  |
| --- | --- |
| **Abstract:** | In this contribution, we would like to conduct all the types of data imperfection that we may encounter in real medical data mining techniques. Then, we propose an algorithm to deal with these information elements in a unified mathematical framework based mainly on possibility and evidence theory. Afterward, we introduce some applications in classification, clustering, diseases diagnostics, case based reasoning, etc. applied on real digestive databases in some hospitals in France under the supervision of medical experts to validate the proposed methods and algorithms. These techniques could be applied to any other disease database without modifications and are useful in machine learning and in building knowledge datasets as an essential step in artificial intelligent for health. |

1. **Introduction**

Perfection is an objective that has never and will never be accomplished in this world. But despite this cold hard truth, mankind still struggles to find it in their life through effort, dedication, and persistence. This is a fact that can be materialized in every aspects, approach, technique, behaviour, in this life. Nowadays, technology is available to help us to collect data by means of cameras, scanners, satellites, etc. and it is available to help us to save these data through the great advances in storage media, databases, data warehouses, etc. Nevertheless, the advances in data acquisition and storage technology have led to a remarkable growth in datasets and consequently to a high difficulty of the human to analyse or to discover the underlying knowledge or patterns when handling such very large databases. Consequently, man becomes rich in data but poor in knowledge. For this reason, data mining is rapidly gaining attention among researchers in the field of information processing, artificial intelligence, and multimedia databases thanks to its potential in discovering useful new patterns and trends that may push the various researches to new frontiers. Similarity measuring is the corner stone in any mining technique because any system whose goal is to analyse or to organize automatically a set of knowledge or data must use by a way or another a similarity operator to establish the resemblance and the relations that may exist in data. But according to the imperfection principle, information elements can be imperfect, uncertain (imprecise, probabilistic, evidential, possibilistic, ambiguous, etc.), and even inexistent (missing data). Additionally, these information elements can have any type and structure regarding their measuring scale (quantitative, nominal, ordinal, etc.), and the challenge that can encounter the poor human is how to deal with information imperfection, uncertainty and heterogeneity all together in a unified framework in this world of imperfection. Unfortunately, this problem hasn’t been tackled in its integrality in the prior works of the literature, and the humble and simple attempts to deal with the missing data problematic alone, the imprecision in its own, or the heterogeneity of information apart; become insufficient, useless, and even incapable to deal with the real data in its general context. Another important factor has to be taken into account when applying data mining techniques, namely the computation time. In this imperfect world, every thing is fast, and many applications have to be achieved in the real-time as soon as possible. Dealing with huge number of data as is the case of data mining incites us to look for a fast constraint-free algorithm which doesn’t occupy a large space in the memory and which doesn’t require complex computations from the processor. Actually, we look for a simple, robust, human-reasoning-like approach that can tackle the imperfection, the uncertainty and the heterogeneity of information elements at the same time in the same context based on a simple mathematical model capable to build bridges with the other mathematical models, in order to be able to conduct with all possible information processing systems (the scalability). We will see along this paper that the possibility theory is the best model we are looking for through many examples of our possibilistic model applied to several data mining techniques to prove the flexibility mixed with the simplicity of this theory when dealing with information heterogeneity and imperfection.

1. **Data Imperfection**

**2-1- From Data to Information**

The word *data* is the Latin plural of *datum*, the past participle of the verb *dare* (to give), hence data mean givens. In the past, the terms *givens* and *data* have been used interchangeably in geometry, mathematics, engineering, and in many other domains to represent any raw symbols that exist but have no significance or meaning beyond their existence. Such usage was the origin of the term *data* as a concept in computer science used to describe numbers, words, images, measurements, inputs, etc. Giving more meanings and descriptions by means of relational connections can transform data into information. For instance, the data “37” given as a number can be transformed into information by adding some other elements like “37 °*C*” which informs us that this measure represents a temperature, and by telling which type of temperature we describe (human or ambient temperature for example). Consequently, information can be defined as a function called the informative function described by means of a model that maps the information definition set that represents the object of our description to the information content set that represents the manner used to describe the information (see figure 1). The notion of the model is very essential to determine the confidence degree of the information element and to understand it. For example, the interpretation of the zero gray level in a resolution cell in the domain of remote sensing imaging in terms of its thematic contents can not be conducted if the physical model of the sensor is not known. If the used sensor is imaging radar, then, the resolution cell contents can be interpreted as being a flat surface or may correspond to a shadow area. On the other hand, if the imaging sensor corresponds to a given spectral band in multispectral imaging, then, the resolution cell thematic content is the content absorbing the emitted electromagnetic signal in the considered spectral band. Here are some more examples of information elements:

* + “The observed lesion **could be** a malign tumor with a **degree of confidence of 0.8**”.
  + “The observed lesion **is** a **benign** tumor”.
  + “The observed lesion is **one of the classes**  **with a belief of 0.9**”.

As we see, in the diagnostic reasoning, information can be described by at least the following parameters:

1. an object or class of objects that refers to the considered patient;
2. a characteristic (attribute) describing the object or the class of objects;
3. the degree of confidence associated with the pair {object(s), attribute(s)} according to the model of this process.

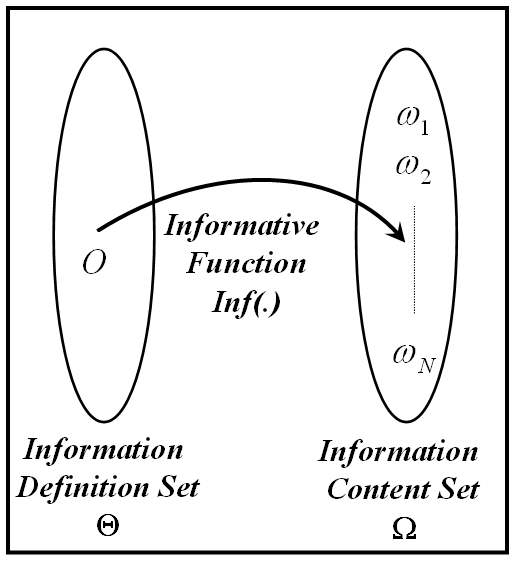


Figure 1. General scheme of information structure (The set of all the subsets of  is called the power set of  and is denoted or  since its cardinality is equal to ).

Let us now illustrate some important examples of information elements using the general scheme of pattern recognition system shown in figure 2.

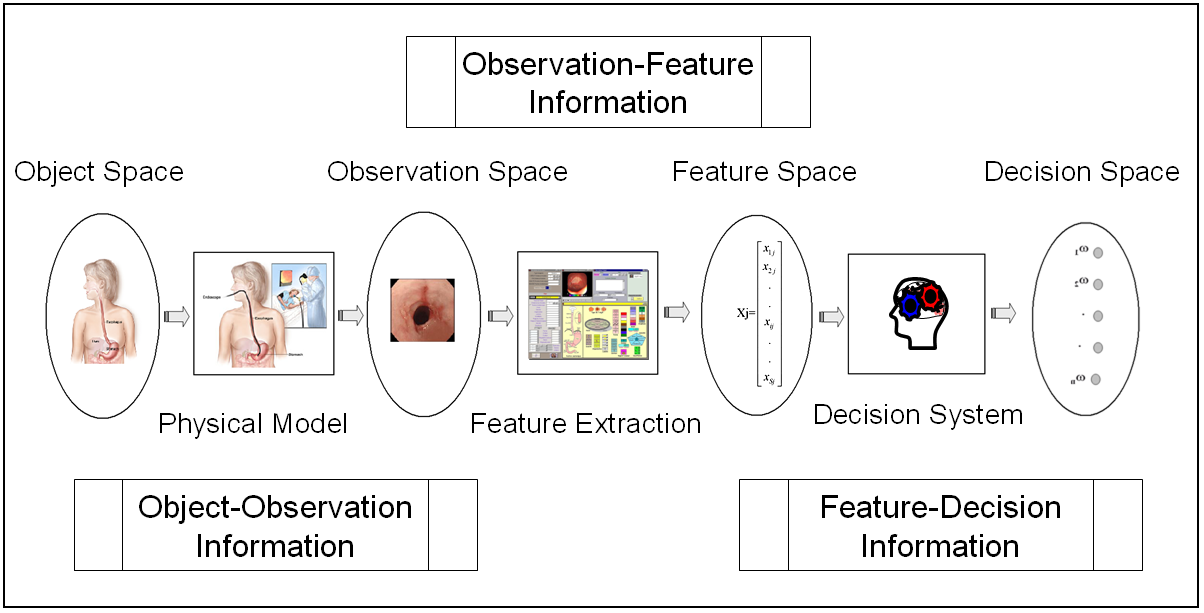


Figure 2. General scheme of pattern recognition systems.

From this scheme, we can distinguish at least the following potential information elements:

1. Object-Observation Information Element: the purpose of the model here is to extract particular facet(s) of real world objects.
2. Observation-Feature Information Element: this information generally aims at reducing the dimensionality of the observation space while maintaining observation representativeness. The model in this information element is mainly related to the concepts exploited by feature extraction techniques.
3. Feature-Decision Information Element: the aim of information here is to relate features to decision, and this is the most common element in pattern recognition systems.

**2-2- Information Characterization**

Any information element can be characterized by means of three main indicators: the exhaustiveness, the exclusiveness, and the confidence degree (linkage strength). We say that the information is exhaustive if , we can find an element  in  such that . This means that  contains all the possible hypotheses, classes, decisions labels, description, etc. This property can also be referred to as “closed world assumption”. We consider that the information is exclusive if there is only and only one element  in  such that . In other terms, there is only one informational content (decision, hypothesis, .etc) associated to the elements of the information definition set. In this case, we say that we have precise information. Concerning the third property, information can have either binary linkage strength (the linkage  is completely true or completely false) or partial linkage strength (the linkage  is associated by a weight, a membership degree, etc.). For example, we need a binary linkage to decide if the cells of a lesion described by pixel grey scale values are benign or malign, whereas we may need a partial linkage to describe the decision concerning the disease of a patient since the latter can have more than one disease at the same time or since the observed symptoms prompt the doctor to attribute membership degrees to some probable pathologies.

**2-3- Information Uncertainty and Imperfection**

There are two key types of challenges encountered in the information structure schematized in figure 1. The first one concerns the uncertainty of information, and is associated to the belief in the veracity of this information by characterizing its conformity to the reality. In fact, the lack of reliability in information sources is a main reason of uncertainty. For instance, the information concerning the variable “temperature” can be uncertain if we have two thermal sensors that provide us with two different numeric values of this variable: 24 °*C* and 26 °*C.* Another type of uncertainty can be ensued on account of the preprocessing steps or owing to the physical model. For instance, the machines installed in care clinic rooms provide us with many types of data and measures as the physiologic signals and the numeric measures estimated from these signals (like estimating the cardiac frequency from the ECG signals for example). The estimation of these measures could be achieved automatically in the machine in many cases, and the access to the algorithms is unavailable for the doctor, besides, these build-in algorithms could be different for different machine marks. The hidden variables are also a major source of uncertainty since many factors and features cannot be quantifiable or qualificatory, or since the relation between the parameters is sometimes indescribable or indefinable. Doctors, when entering the patient’s room, can have impression and intuition of patient’s state by means of the color of skin, the smell, etc. Actually, the uncertainty of data is a delicate widespread problem in medicine since 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.

The second challenge in information processing is the imperfection of data. Unlike the uncertainty which is associated to information source reliability, the imperfection is always related to the content of the information (information content set in figure 1). In this context, we have the following types of imperfect information:

**2-4- Imprecise Information**

We say that the information is imprecise if it is exhaustive, exclusive, has a total confidence degree and our knowledge concerning the identity of  in (), is described via a subset  (see figure 3). The following are examples of imprecise information:

1. “The observed lesion is **one of the classes** ” (qualitative imprecision);
2. “The diameter of the lumen of esophagus **is between** 3 and 4 centimeter” (quantitative imprecision).

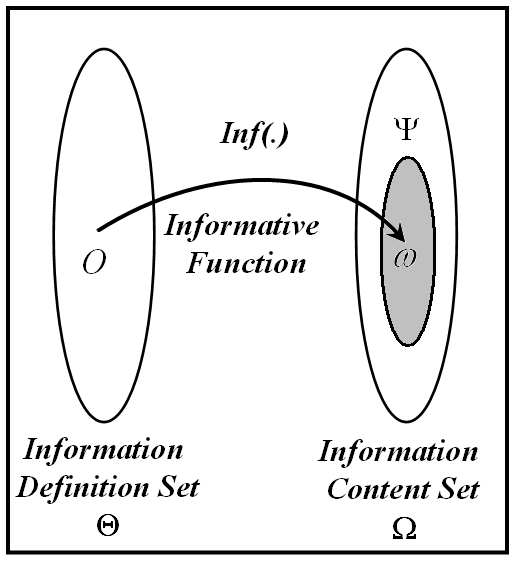


Figure 3. Imprecise information

**2-4- Probabilistic Information**

We say that the information is probabilistic if it is exhaustive, exclusive, has a total confidence degree and if our knowledge concerning the identity of  in (), is described via a probability density function (p.d.f) defined as:,  (see figure 4). Two examples of probabilistic information are given below:

1. “The observed tumor **is** **highly probable** to be a **benign** tumor ()”, In this case we know also that ();
2. “We can’t decide whether the tumor is benign or malign” (total ignorance). In this case, if we want to model our knowledge concerning the classification of the object by a probability distribution we can assume that: ().

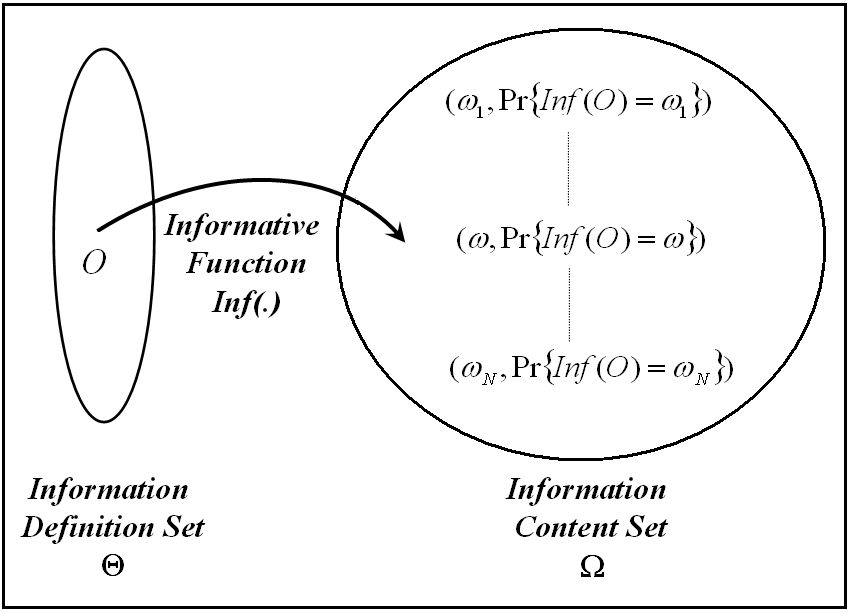


Figure 4. Probabilistic information

**2-5- Evidential Information**

We say that an information element is evidential if it is exhaustive, exclusive, and has a total confidence degree and if our knowledge concerning the identity of  in (),  is described via basic belief assignments (belief functions) defined upon  as:   (see figure 5 that shows the differences between the probabilistic and the evidential approach). The following are examples of evidential information:

1. “The observed lesion is **one of the classes**  **with a belief of 0.9**”.
2. “We can’t decide whether the tumor is benign or malign” (total ignorance). In this case, if we want to model our knowledge concerning the classification of the object by the belief masses, we can assume that: .

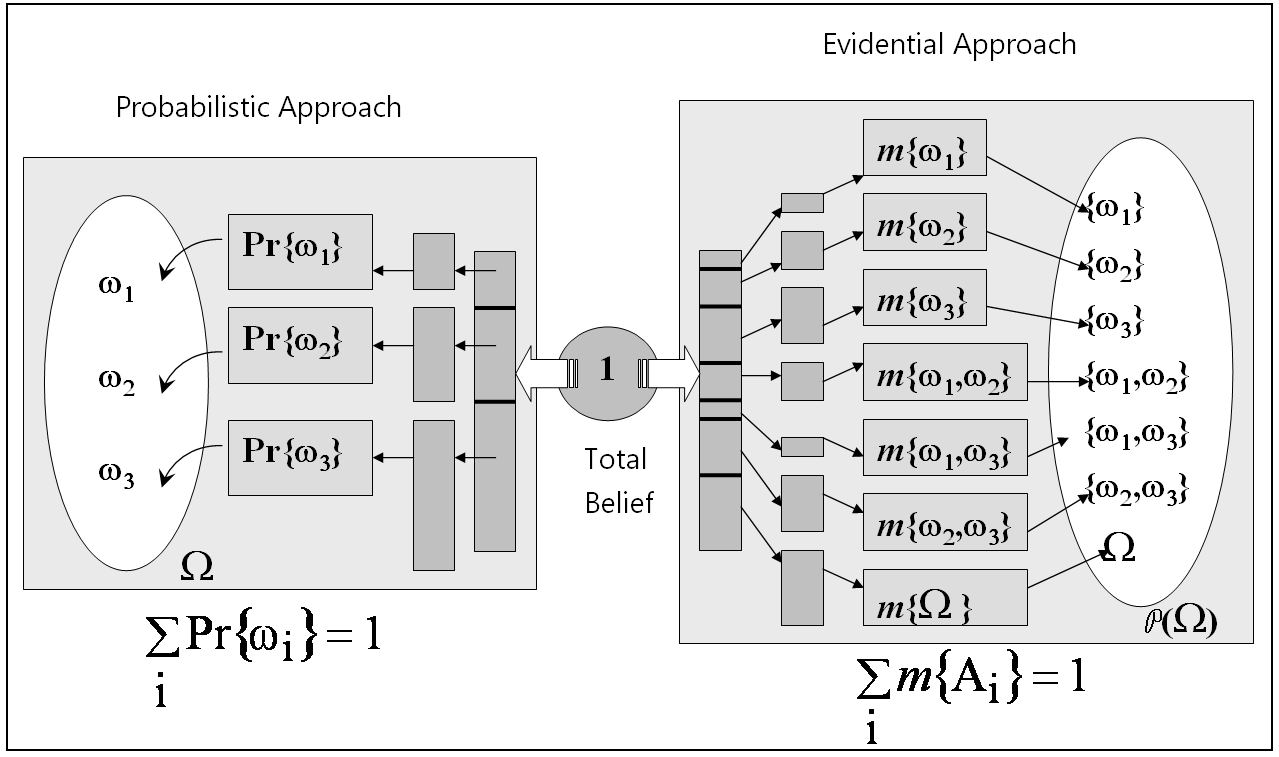


Figure 5. The main difference between the probabilistic and the evidential approaches.

**2-6- Ambiguous Information**

Ambiguous information is information characterized by ill-defined attributes, or by some classes having ill-defined limits (see figure 6). Instead of defining the membership of a known object in an ill-known class by a strict relation (“1” as belong to, “0” as doesn’t belong to), we consider that the membership can be fuzzy. Consequently, we define a membership measure of an object in some class (which has ill-defined limits). Here are some examples of vague information:

1. “The observed lesion is somehow heterogeneous”;
2. “The observed lumen of esophagus has a very big diameter”.

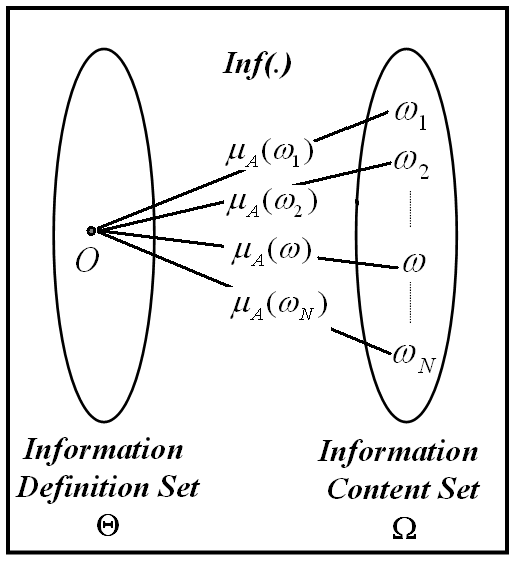


Figure 6. Ambiguous information where *A* represents a fuzzy set defined upon  and generated by , and  represents the association strength of .

**2-7- Possibilistic Information**

When the knowledge about the informational content of an exhaustive and exclusive information element is ambiguous, this element is characterized as a possibilistic information element. So, one and only one class  (hypothesis, decision, label, etc.) from the information content set  is associated with the object , however; we have an ambiguous knowledge about this :

1. “we know that  is peter, , and peter is young”.
2. “ is an object, , and this object is close to ”.

As we have seen information content set data can have diverse formats and can be stored using a variety of different storage modes. At the most elementary level, a single unit of information is a **value** of a **feature** (attribute), where each feature can take a number of different values. The **objects (records)**, described by features, are combined to form data sets, which in turn are stored as flat files and in other formats using databases and data warehouses in order to extract potential valuable patterns or rules (knowledge) using the **data mining** **techniques**. The relationships among the above concepts are depicted in figure 7.

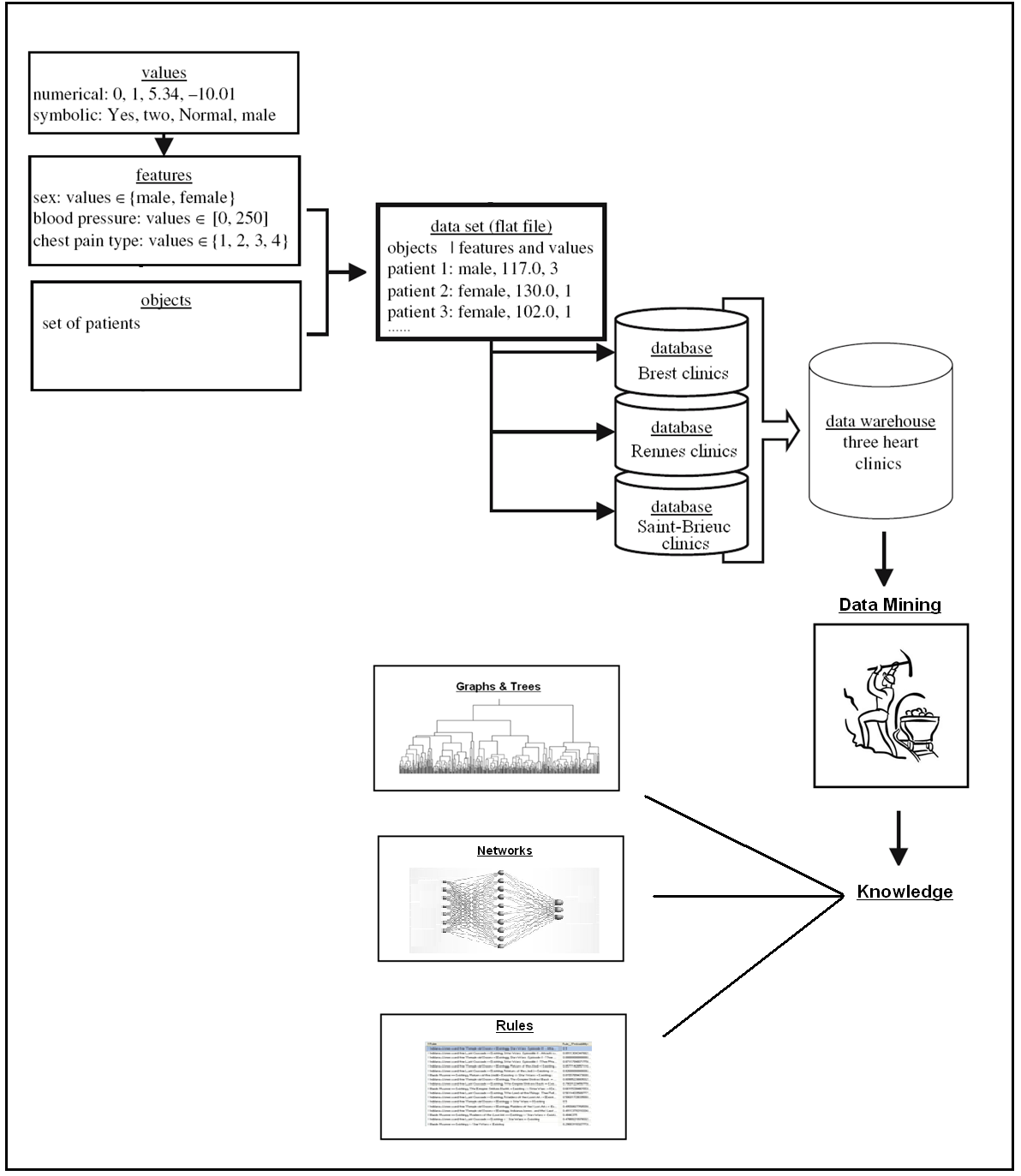


Figure 7. Relationships between values, features, objects, data sets, databases and data warehouses.

Actually, there are two main types of values: quantitative (numerical) and qualitative (symbolic). Numerical values are expressed by numbers, for instance, real numbers (-1.09, 123.9), integers (1, 44, 125), etc. In contrast, symbolic values usually describe qualitative concepts such as color (white, red, etc.) or sizes (small, medium, etc.).

Features (also known as attributes) are usually described by a set of corresponding values. For instance, height is usually expressed as a set of real numbers. Features described by both numerical and symbolic values can be either discrete (categorical) or continuous. Discrete features concern a situation in which the total number of values is relatively small (finite), while with continuous features the total number of values is very large (infinite) and covers a specific interval (range).

A special case of a discrete feature is the binary (dichotomous) feature, for which there are only two distinct values. A nominal (polytomous) feature implies that there is no natural ordering among its values, while an ordinal feature implies that some ordering exists. The values for a given feature can be organized as sets, vectors, or arrays. This categorization of data is important for practical reasons. For instance, some preprocessing and data mining methods are only applicable to data described by discrete features. In those cases a process called discretization becomes a necessary preprocessing step to transform continuous features into discrete ones, and this step must be completed before the data mining step is performed. Objects (also known as records, examples, units, cases, individuals, stimuli, data points) represent entities described by one or more features. The term multivariate data refers to situation in which an object is described by many features, while with univariate data, a single feature describes an object.

Let us consider an example concerning patients at a heart disease clinic. A patient is an object that can be described by a number of features, such as name, sex, age, diagnostic test results such as blood pressure, cholesterol level, and qualitative evaluations like chest pain and its severity type. An example of a “patient” object is shown in Figure 8:

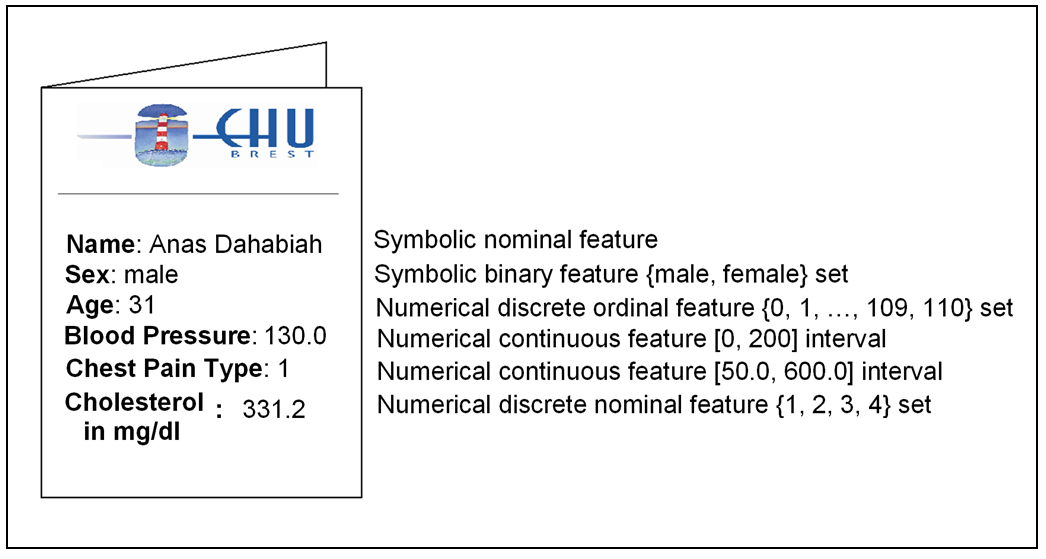


Figure 8. An example of an object (patient record)

In this example, the cholesterol has been given as a numerical value, however; owing to the limited comprehension of numbers by humans, who are the ultimate user of the information, this value can be expressed in terms of aggregated information such as “high” or “low” level of cholesterol, or according to the final objective or according to the mathematical model that we build, this value can be represented by an interval (figure 9-b), or by using the notions of the fuzzy sets (figure 9-c) or the rough sets (figure 1-9-d).

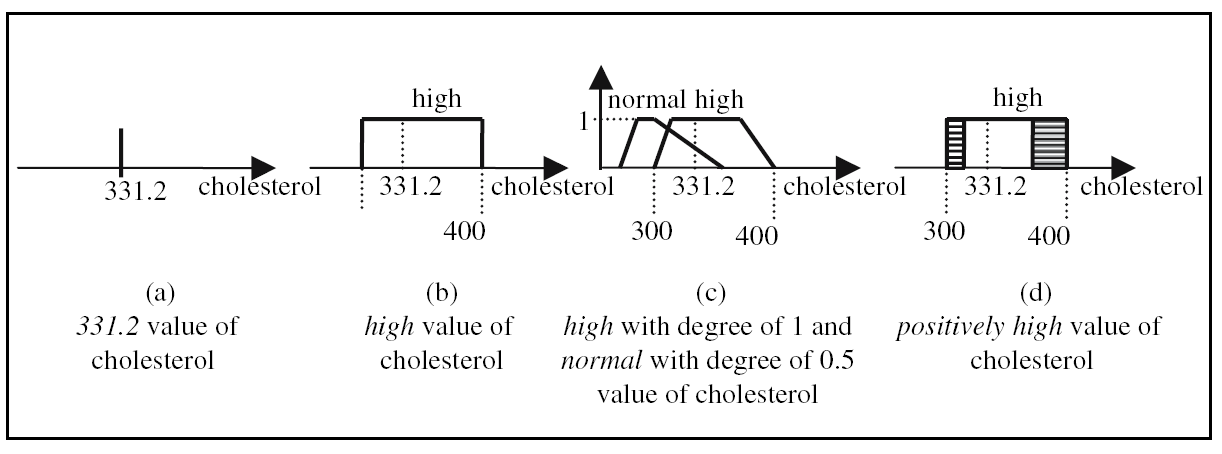


Figure 9. Several data types of the same attribute.

The objects described by the features are grouped to form data sets. Many data mining and statistical data analysis tools assume that data sets are organized as flat files (a rectangular formatted table composed of rows and columns, where the rows represent the objects and the columns represent the features). Flat files are used to store data in a simple text file format, and they are often generated from data stored in other more complicated formats, such as spreadsheets or databases. Now, in order to analyse (to mine) the data of the heart clinics that are coming from different cities, a data warehouse is build to be a repository of the collected data by applying a set of processing steps (data loading, uploading, cleaning, transformation, and integration).

1. **Data Mining and Knowledge Discovery in Medical Databases (KDD)**

Nowadays, technology is available to help us to collect (thanks to scanners, satellites, cameras, sensors, etc.) and to store (thanks to databases, data warehouses, and the variety of repositories) all the types of data that we can imagine. Consequently, the amount of data kept in computer files and databases is growing at a phenomenal rate. On the other hand, the users and the owners of these data who are expecting more sophisticated information and knowledge of them, are confronted with the challenge to cope with the flood of the generated data, and are starving for interpreting these data in search for new knowledge. Data mining techniques step in here to fulfill these needs, by exploring and analysing the database with the aim of extracting new unknown potential knowledge. In data mining, we are not trying to find a needle in a haystack because database management systems know very well how to do that. In the contrast, we are merely trying to understand the consequences of the presence of the needle, if it exists. Data mining can be defined as the set of data exploration techniques that permits to extract from a large database interesting potential knowledge in the form of description models (rules, regularities, trends, constraints, etc.) with the aim of describing the actual behavior or predicting the future behavior of the data. According to this definition, the main task of data mining is to fit a model to the data by choosing the type of the model (e.g. classification), and then by choosing the appropriate method to construct this model (e.g. using the decision trees to classify the data). Data mining models can be either predictive or descriptive in nature. A predictive model makes a prediction about values of data using known results found from different data. Predictive modeling may be made based on the use of other historical data. For instance, giving some types of medicines to a patient might be refused, not because of the patient’s own history, but because the current situation of the patient is very similar to that of other patients who tried these medicines without any improvement. Predictive model data mining tasks include classification, regression, and prediction. Unlike the predictive model, a descriptive model serves as a way to explore the properties of the data examined, not to predict new properties but identify patterns and relationships in data. Clustering, summarization, and association rules are usually viewed as descriptive in nature. Unfortunately, the terms “*knowledge discovery in databases (KDD)*” and “*data mining*” are being used interchangeably in this field, although the data mining process represents just a step in knowledge discovery (see figure 7). So if we imagine that the knowledge discovery process is a vehicle, then data mining process is the motor that drives it forward. Another misusage of the terms arises when we use the words data, information, and knowledge interchangeably. The main purpose of the process is to extract the knowledge not the information. In fact, knowledge is the appropriate collection of the amassed information in a certain context. For example, elementary school children memorize, or amass knowledge of the “times table”. They can tell you that “” because they have amassed that knowledge (it is included in the times table). But when they are asked to give the result of “”, they can not respond correctly because that entry is not in their times table. To correctly answer such a question, a true cognitive and analytical ability is required. And that leads us to the next level “the understanding” (see figure 10). The extracted Knowledge in any KDD process can be represented by several form, like as ***rules*** (if “condition” then “conclusion (action)”), ***graphs*** (in this case, the concepts of the knowledge are presented as nodes in the graph, and the linkages between the concepts (associations, dependencies, etc.) are presented as the edges of the graph.), or as ***networks*** (the networks can be regarded as generalized graphs in the sense that at each node of the graph we encounter some local processing capability. In other words, the network doesn’t only represent the knowledge itself, but it also contains the underlying processing realized at the local level).

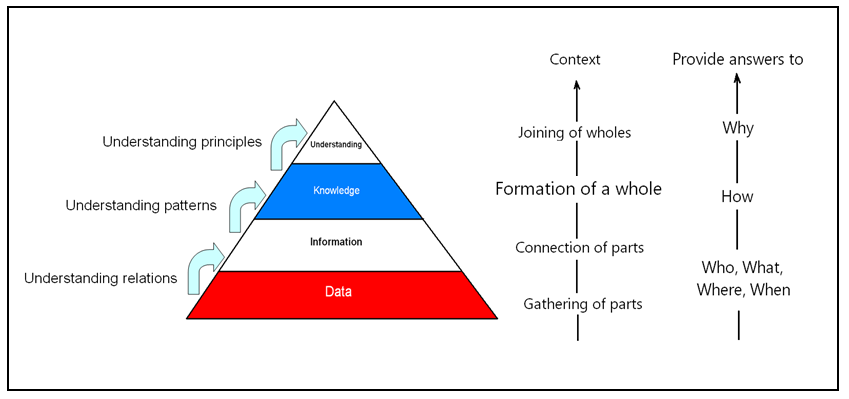


Figure 10. Data, information, and knowledge relationships.

**Examples of Basic Data Mining Tasks**

* + **Classification**

Based on the characteristics of an object, classification maps this object into predefined group or class (see figures 1 and 2, supposing that  is the set of all the possible classes). It is often referred to as supervised learning because the classes are determined before examining the data. We note here that ***pattern recognition*** is a type of classification where an input pattern is classified into one of several classes based on its similarity to these predefined classes.

* + **Clustering**

Clustering is similar to classification except that the groups are not predefined, but rather defined by the data alone. Clustering is alternatively referred to as unsupervised learning or segmentation. It can be thought of as partitioning or segmenting the data into homogeneous groups that might or might not be disjointed (we will see later). The clustering is usually accomplished by determining the similarity among the data on predefined attributes. The most similar objects gather together in the same cluster.

* + **Prediction**

Based on past and current data, prediction algorithms involve in predicting future data states. As a matter of fact, prediction can be viewed as a type of classification. The main difference is that prediction predicts a future state rather than a current state, and here we refer to a type of application rather than to a type of data mining modeling approach.

* + **Summarization**

Summarization (also called ***characterization*** or ***generalization***) maps data into subsets with associated simple descriptions. It extracts or derives representative information about the data with the aim of succinctly characterizing the contents of the databases basing on the similarity among their elements.

* + **Association Rules**

*Link analysis*, alternatively referred to as *affinity analysis* or *association*, refers to the data mining task that aims to discover potential relationships among the attributes. The best example of this type of application is to determine association rules. An *association rule* is a model that identifies specific types of data associations depending on the similar distinct characteristics of the object in a given group like if the patient’s cholesterol is so or so and his age is more than *x* , he has such or such disease.

1. **Proposed algorithm to measure similarity of patient records**

Any mining system whose goal is to analyse or to organize automatically a set of data or knowledge must use, by a way or another, a similarity operator to estimate the relations and the resemblance that exist in the examined data. Looking closely, for instance, to the definitions of data mining tasks presented before or to other possible tasks (as outlier detection, time series analysis, etc.) presented in the literature, or to the mechanisms of other information processing techniques (as content-based case retrieval, case-based reasoning, etc.) proves the importance of similarity measuring and shows the need to use an appropriate measure. However, as we have already shown, the amount of data is increasing rapidly, with all possible data types (continuous or discrete), and with all possible data measuring scale (quantitative, qualitative, binary, distributions, etc.). Accordingly, the demanded similarity measure has to take all the states of data into account, providing a rapid processing computation time because we are handling huge amounts of data. In the following, we are presenting a list of commonly used measures, supported by numeric examples and applications when necessary, to measure the similarity between two vector (objects)  and  consisting of  distinct variables (attributes, features, stimuli, observations, examples, characteristics) represented as: , . (T is the transpose of a vector,   is the *ith* attribute of  *where *). We suppose that the data set consists of *n* objects. We will use the terms similarity (proximity), dissimilarity (distance) interchangeably, since they present two faces for the same coin (we can get one from the other by applying a suitable decreasing function). The distance between the variables will be denoted as *d* (small letter), whereas the distance between classes (or clusters) will be denoted as *D* (capital letter). The list that we present is not exhaustive because we need books to accumulate all the proposed similarity measures, but it is very useful, general, and helpful for lots of researchers and PhD. students in many disciplines.

Traditional similarity (dissimilarity) measures (Minkowski, Canberra, Hamming, Jaccard, etc.) suppose generally that the value of each attribute is precise (disregarding the existence of imprecise data), certain (ignoring the existence of uncertain values), and given (neglecting the existence of missing values) while on the contrary, real databases contain a remarkable amount of incomplete, imperfect, or even uncertain values. In the literature and previous works, information imperfection has been treated superficially in measuring the similarity, and a few humble attempts have taken some of its aspects like the imprecision into account using complicated approaches that depend on several parameters whose estimation is empirical, complicate, and uncertain in its own.

Moreover, 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 cannot 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 easily modeled and combined in the possibility theory framework that we propose as we will see later.

In addition, some constraints and conditions should be considered when dealing with each similarity measure. For instance, 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 the interval of some measures like the angular separation similarity and Pearson correlation coefficient belongs to [-1,1]). Besides, the meaning, the interpretation, and the objectives of the similarity measure must be taken into consideration. For example, we ought to keep in mind that the angular separation similarity and Pearson correlation-coefficient-based distance tends to disclose the differences in shapes rather than to detect the magnitude of differences. Furthermore, The scale of measurement of the attributes is an important consideration when using some distances like the Euclidean distance measure. Changing the scale can affect the relative distances among the objects. For example, suppose three objects have the following bivariate measurements. Using the Euclidean distance , the dissimilarity matrix of these objects  will be given as:

.

However, if we multiply  by 100 (as, for example, in changing from meters to centimetres), the matrix becomes:

,

and the largest distance is now  instead of  (the distance rankings have been altered by scaling).

To counter this problem, each variable could be standardized in the usual way by subtracting the mean and dividing by the standard deviation of the variable. However, such scaling would be ordinarily based on the entire dataset, i.e. on all the S values in each column of X. In this case, the variables that best separate clusters in a clustering problematic might no longer do so after division by standard deviations that include inter-cluster variation.

In addition to the aforementioned constraints and limitations, many implementation caveats related to some technical and mathematical problems must be considered. Table 1 contains some examples.

Table 1. Some examples of the mathematical cautions taken into account when measuring the similarity.

|  |  |
| --- | --- |
| Problems | Examples |
| Division by zero | Kulczynski , Pearson χ2 , Neyman  χ2 , Kullback-Leibler, Jeffreys, Taneja , and Kumar-Johnson. |
| 0/0 | Canberra, Wave Hedges, Harmonic  mean, Squared χ2 , Probabilistic Symmetric χ2 , Divergence , Clark , and Additive Symmetric χ2. |
| Log of 0 | Jeffreys |
| 0log0 | Kullback-Leibler, K divergence, Topsoe, Jensen-Shannon, Jensen difference, and Taneja. |
| The calculation of the matrix inverse may cause some computational burden in large scale data. | Mahalanobis distance. |

As we see from the above discussion, the traditional measures are overburdened with a lot of conditions and constraints that must be carefully considered.

* + **Probabilistic Uncertainty vs. Ambiguity**

There are basically two main models concerning the imperfection of information. The first one, called the *probabilistic uncertainty*, aims to discover the unknown information content , for a given object *O* from (figure 1), given that the information are characterized by the exhaustiveness (closed world assumption), the exclusiveness, and the binary linkage strength. For instance, casting a dice and hiding the result is a example of the probabilistic uncertainty imperfection type, since we can obtain merely ***one*** result included ***inside*** the reference set (information content set) , and the linkage strength is ***binary*** here, since when I get any value belonging to the content set, the linkage of this result is characterized either as true or as false. In this context, there are three essential well-known theories: probability theory, evidence theory, and possibility theory. The basic objective of these theories is looking for the unknown informational content . In the second model, called the ambiguity, the information elements can be described according to the three indicators that characterize any information element as follows: concerning the exhaustiveness, information elements are always processed in the closed world assumption (all the possible hypotheses are known). Regarding the exclusiveness, information elements can be non-exclusive, as the case with a patient from  who has two different pathologies (that can take place simultaneously) from the list of the all known imaginable pathologies  (remind that putting all our belief in the occurrence of an hypothesis in the probabilistic uncertainty model means that there is no chance or belief assessed to the other hypotheses, since the sum of all the probabilities has to be set to “1”). Figure 11 depicts an example of a patient with two different pathologies (spot and ulcer) occurred at the same time. Unlike the probabilistic uncertainty, in the ambiguity model an information element can be associated with several informational contents with weights or membership degrees (partial linkage strength). The objective here is to combine or to characterize the available ambiguous information at one time as a fundamental step in decision reasoning or information fusion. This can be done by means of the tools and approaches proposed in the framework of the fuzzy set theory proposed by Zadeh. As an example, let us consider the case of a doctor who is looking for an efficient and a cheap drug at one time. Actually, all the drugs are more or less efficient, more or less cheap, and these two criteria can be characterized by two membership restrictions that show to which degree each drag correspond to the category under consideration. To combine and to take account of these two criteria, we can calculate the minimum of the corresponding membership degrees estimated from their function. Thus, the objective is here to combine the criteria rather than to find the only content information element.

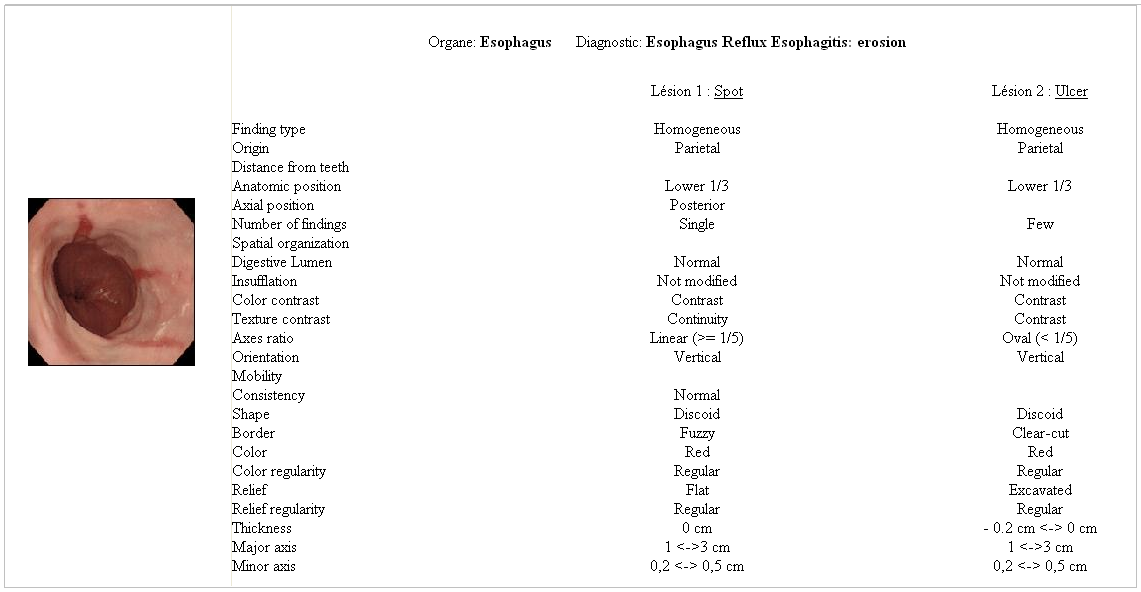


Figure 11. An example of an ambiguity model.

The clear distinction between these two models (the probabilistic uncertainty and the ambiguity) is very essential in information processing when choosing the appropriate method and the suitable theory to deal with the given information elements and to avoid the common errors widely committed in the literature concerning their description and their characterization. For example, if a patient has two pathologies (two informational contents  and ) that take place at the same time, we must get use of the fuzzy set theory when describing and characterizing this ambiguous information, whereas if the doctor believes that the patient has only one pathology but there is some hesitation between two of them, let us say between  and , we must get use of the evidence theory for instance to handle this imprecise information element () that belongs to the probabilistic uncertainty model. Figure 12 depicts the difference between these two cases.

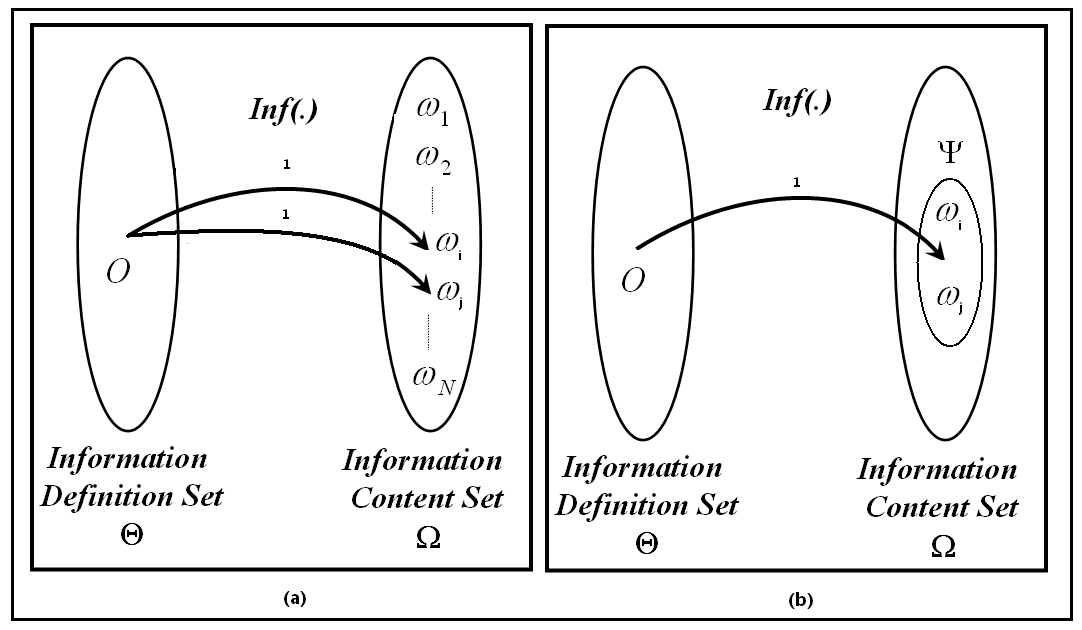


Figure 12. The main differences between (a) ambiguous and (b) imprecise information.

* + **Evidence Theory**

Evidence theory has been found to deal with the probabilistic uncertainty (the exhaustive and exclusive information elements) when the available knowledge can not allow to distinguish between several informational contents due to the confusion, the ignorance, or/and the incompleteness of information that requires another evidence or observation. This type of imperfection, namely the imprecision, has not been taken into account in probability theory, and any attempt to handle it using the approaches and the concepts provided by probability theory can lead to undesirable, illogical, and conflicting results. For example, supposing that the informational contents of an object  (figure 1) are contained in a finite unordered set called the frame of discernment, defined as  for example, and assuming that the distinction between  and  is impossible due to the physical model used as the informative function. This can take place for instance when a medical image technique (like X-ray, IRM, or scanner) is unable to distinguish between two different types of tissues or organs like  and , as a consequence of the incapability of a radar to discriminate between  and  in the remote sensing images, or , as a third example, because of the hesitation of a doctor to choose an alternative between two drugs, two diagnosis, etc. on the light of the available evidence that provides him with a partial knowledge. The main question that arises here is how to exploit an impression like  in a probabilistic information context. Putting all our belief in  alone or in  alone in probability theory is irrational approach. Dividing our belief between the two hypotheses in such a way that () cannot interpret the model incapability to discriminate these two informational contents. The only rational and intuitive approach is to put all our belief in the composed hypothesis  rather than in the elementary hypotheses (singletons)  or .

Evidence theory permits to allocate the total belief normalized to the unity value over all the possible subsets of the information content set (see figure 5), through what is called the basic belief assignment denoted as *m* or *BBA*, defined as:

****

****

**; **

The subsets (*A* of ) for which  are called the focal sets of *m*. Each focal set *A* is a set of possible contents of  and the number  can be interpreted as a fraction of a unite mass of belief, which is allocated to *A* on the basis of a given evidential corpus. Complete ignorance corresponds to put all our belief in the information content set  (), and the perfect knowledge of the content of  is represented by the allocation of the whole mass of belief to a unique singleton of  (*m* is then called a certain *BBA*). Another particular case happens when all focal sets of *m* are singletons. In this situation, *m* is equivalent to a probability function and is called a Bayesian *BBA*. A basic belief assignment *m* can be equivalently represented through two non-additive fuzzy measures; the belief function (*Bel*) and the plausibility function (*Pl*). The belief function of an event *A* for which () is the minimum belief mass that can be associated to the realization of this event, whereas the plausibility function of an event *A* for which () is the maximum belief mass that can be associated to the realization of this event (see figure 13).



 where  is the complement set of *A*.

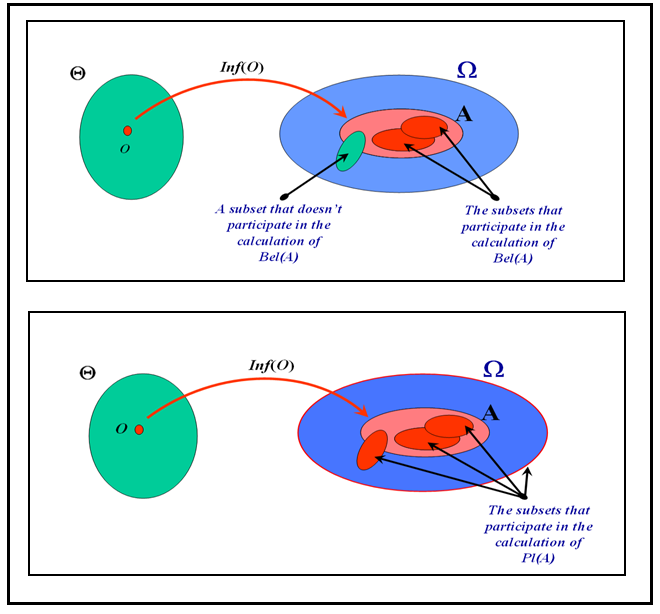


Figure 13. The belief and plausibility functions.

The functions of belief and plausibility respectively represent the inferior and the superior bounds (limits) of the probability allocated to an event . These two functions boil down to a unique probability measure when *m* is a Bayesian *BBA*. Consequently, any probability distribution that realizes ( for ) is said to be compatible with the given mass distribution. The pignistic probability  (where stands for the cardinality of the subset *A*) is a very well-known example of such compatible distribution.

For a given distribution of belief masses, and supposing that the uncertainty has a probabilistic nature, the uncertainty of the occurrence of an event  can be qualified by the confidence interval () and the uncertainty degree of the value of the probability of *A* ().

* + **Evidential Information Fusion**

One of the most well-known applications of evidence theory is information fusion. Combining two *BBAs*  and  representing distinct items of evidence coming from two different sources concerning the content of  could be achieved using any fusion rule proposed in the literature. We give in the following three examples of three very widely-used rules:

* *Dempster-Shafer rule*: , where  is called the conflict degree between  and . It may be seen as the degree of disagreement between the two information sources.
* *Yager rule*: the total conflicting mass is completely assigned to the information content set :  , ; and .
* *Dubois-Prade rule*: when two belief sources support two focal elements in contradiction (having an empty intersection), we put our belief in one of the focal elements instead of . .
* **Evidential Reasoning and Decision Making**

Base on the basic belief assignments, the essential evidence tools like the belief and the plausibility functions, and after combining the information coming from different sources, one and only one decision () has to be adapted (probabilistic uncertainty). In the following, we present three well-known criteria used to help to take this decision:

* Maximum plausibility criterion (optimistic solution): the decision can be taken based on the superior limit of probability of the singletons under consideration, since the maximum plausibility assures a minimum conflict (.
* Maximum belief criterion (pessimistic solution): the decision can be taken based on the inferior limit of probability of the singletons under consideration. This means that we intend to choose the singletons whose belief mass is the greatest. (.
* Maximum pignistic probability: as a compromised solution between the optimistic and the pessimistic criterion, we can consider the maximum of the pignistic probability as: (.
* **Possibility Theory**

Possibility theory deals with the probabilistic uncertainty imperfection in which the content of an object  is represented via an ambiguous knowledge instead of a probability distribution, in such a way that we can get use of all the powerful tools and approaches introduced in the fuzzy set theory to handle the ambiguous information and by taking account of non-probabilistic uncertainties of the event at the same time, with the objective to assess to what extent the occurrence of an event is possible and to what extent we are certain of its occurrence (realization), without, however; knowing the evaluation of the probability of this occurrence. This can happen, for instance, when there is no similar event to be referred to.

Possibilistic knowledge can be represented either via a possibility distribution that interprets the state of knowledge at the singleton level (possibility degree is given to each informational content of  due to the incapability to know the only  ) or via the possibility and the necessity measures that reflects the state of knowledge at the event level (composed hypothesis level) as in the evidence theory. This type of representation enables us to deal with probabilistic uncertainty in an ambiguous environment, by applying the fuzzy set theory tools in modeling the probabilistic problems.

* **Possibility Distribution**

A possibility distribution is a mapping  that assigns a coefficient between “0” and “1” to each informational content, satisfying the normalization condition: . In this context, we have two extreme forms of knowledge:

* *Complete knowledge*: for the informational content ,  and for .
* *Complete ignorance*:   (all the contents of  are possible).

The possibility distribution can be induced by physical constraints when it interprets the degree of ease, efficiency, adequacy, etc. It can be also deduced from the statistics, histograms, etc. by applying Dubois-Prade probability-possibility distribution, or in the case of a possibilistic information defined by a fuzzy restriction characterized by a membership function , the possibility that  is numerically equal to the grade of the membership of  to *A* ().

* **Possibility Measure**

A possibility measure is a mapping  such that:

*  ;
* 

If only two subsets are considered, then  

We can interpret this measure as follows:  represents the extent to which it is possible that the subset (or the event) *A* of  occurs. If  *A* is impossible. If  *A* is absolutely possible. The intersection of two subsets of  cannot be calculated from the possibility measures of these subsets. The only information we obtain from possibility measure definition is that . In other words, two subsets can be individually possible (), but jointly impossible (). We say that a possibility measure  is completely defined if we can assign a possible coefficient to all the singletons of  (in other words, if  is defined). Thus, from a possibility distribution , we construct a possibility measure to any subset of  as follows: . Conversely, from any possibility measure , we construct a possibility distribution as follows:  . Figure 14 depicts four examples of calculating the possibility measure taking account of imprecise and ambiguous possibility distributions and of imprecise and ambiguous subsets .

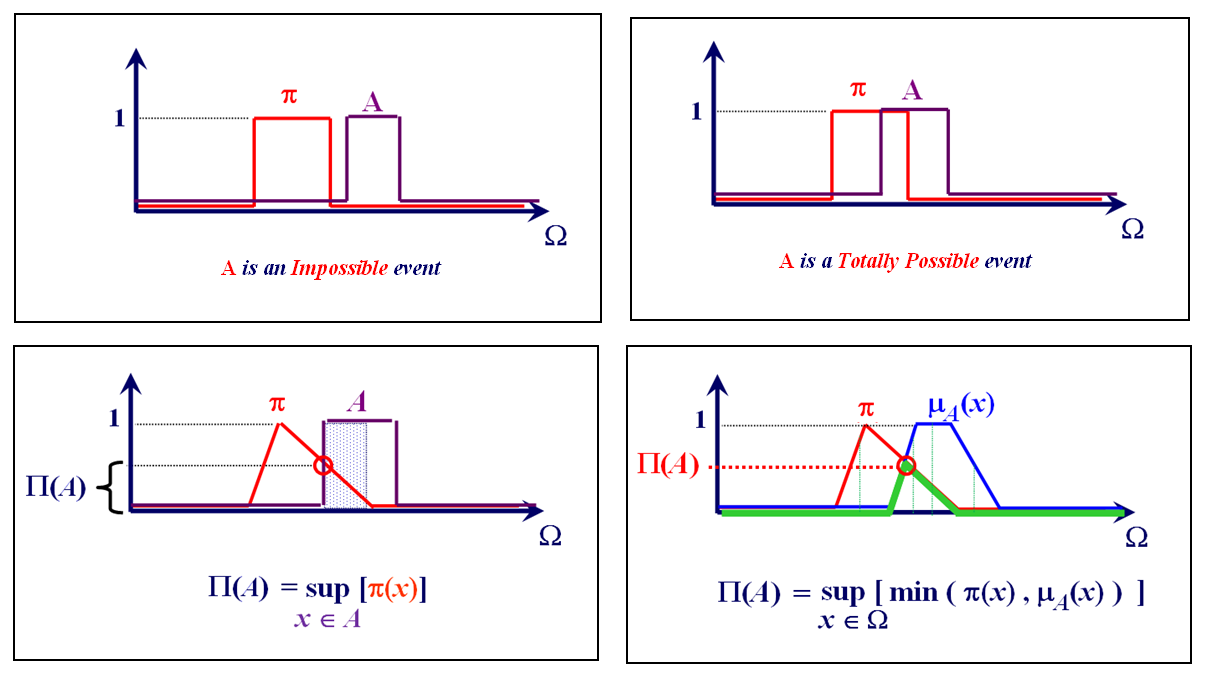


Figure 14. Four general illustrative examples of calculating the possibility measure of an event *A* given a possibility distribution .

In the case of two universes  and , we need to define the extent to which a pair of informational content  is possible with  and . The joint possibility distribution  on the Cartesian product  is defined for any  and  and it expresses the extent to which  and  can occur simultaneously. The global knowledge of  through the joint possibility distribution  provides marginal information on  and  by means of the marginal possibility distribution, for instance for :



Which satisfies:



We remark that a joint possibility distribution provides uniquely determined marginal distributions, but the converse is false. Determining a joint possibility distribution  *on*  from possibility distributions  on  and  on requires information about the relationship between events on  and *.* If we have no information,  cannot be known exactly. The universes  and are *non-interactive* if . This possibility distribution  is the greatest among all those compatible with  and . Two variables respectively defined on these universes are also call non-interactive. The effect of Ω1 on Ω2 can also be represented by means of a conditional possibility distribution πΩ2/Ω1 such that: , , . For a combination operator , generally the minimum or the product. For example, if we consider the universe  of diseases and we add a universe  of symptoms,  is the possibility degree that a patient presents symptom  and  is the possibility degree that a patient suffers from disease. Now, for a disease  and a symptom , we define the possibility degree  that the pair  is possible.  has clearly an influence on  and these universes are interactive. Let us define also the conditional possibility degree  that the symptom is  given that the disease is . For instance, if the available information provides the values  and , then . This means that if the disease  is relatively possible for a given patient and if the symptom  is completely possible when  is present, then it is relatively possible that the given patient presents both disease  and symptom .



* **Necessity Measure**

The possibility measure provides an information on the fact that an event can occur or not, but it is not sufficient to describe the uncertainty about the event and to obtain a conclusion from the available knowledge. For example the event *A* and its complement can be totally possible at the same time (), which means that we have an absolute uncertainty about A (see figure 15).

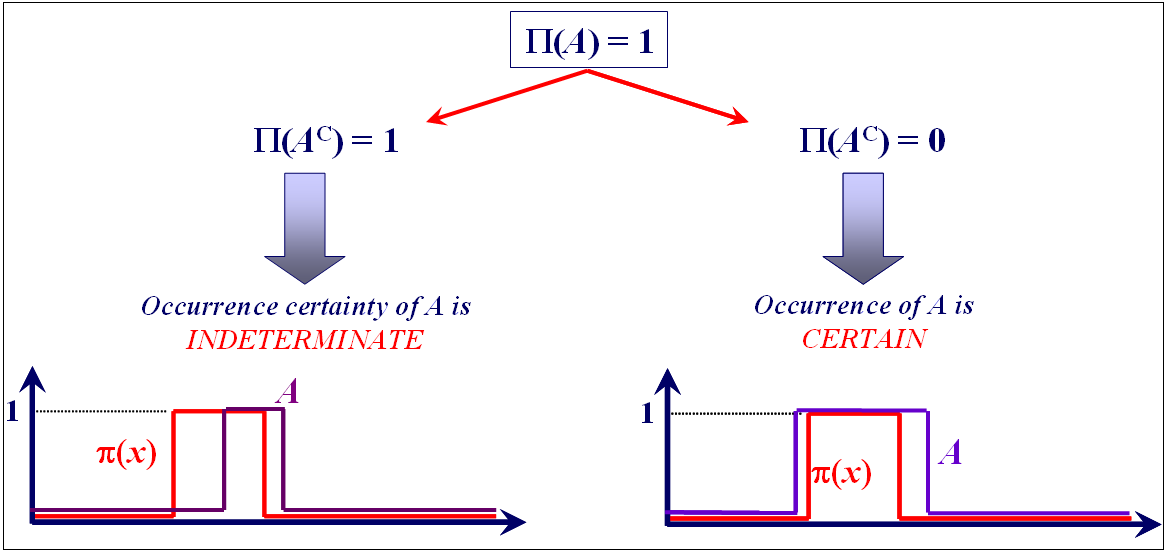


Figure 15. is a good indicator of the occurrence certainty of *A*.

A solution to this problem is to complete the information on *A* by means of the measure of necessity on .

A necessity measure is a mapping  such that:

*  ;
* 

In the case of two subsets of :

We can interpret this measure as follows:  represents the extent to which it is certain that the subset (or the event) *A* of  occurs. When we have no certainty about the occurrence of the event *A*. When  we are absolutely certain that the event *A* occurs (figure 15). The necessity degree of the union of the subsets of  is not known, precisely, but we can know the lower bound: . For the necessity measure, we don’t need a necessity distribution *n* similar to the possibility distribution  in the case of the possibility measure, since there is a duality between the possibility and the necessity measures defined as:  Accordingly, we are certain that *A* occurs (*N(A)=1*) if and only if  is impossible (). The schema depicted in figure 16 shows a general illustrative example.

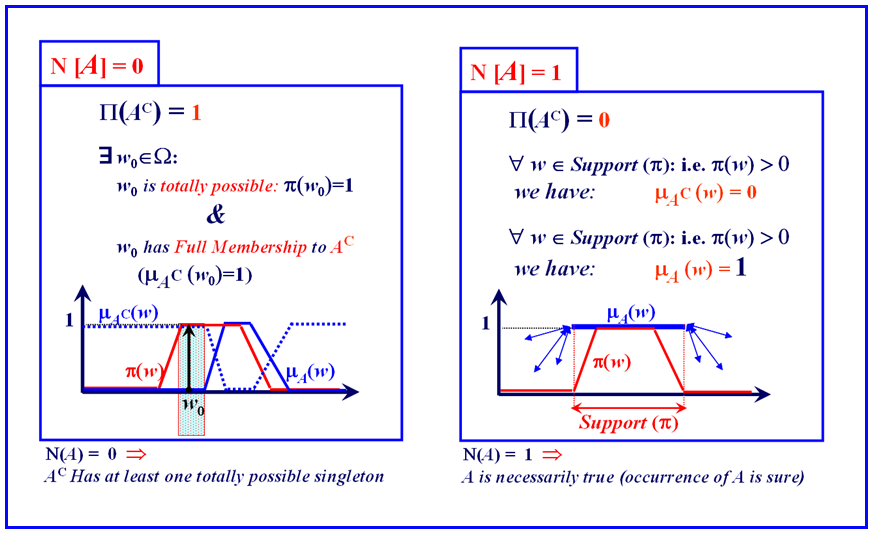


Figure 16. An illustrative necessity measuring schema

If  is defined from a possibility distribution , we can define its dual necessity measure by: , which means that the definition of possibility distribution is sufficient to determine both the possibility and the necessity measures.

* **Possibility and Probability Distribution**

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. In this section, we introduce probability-possibility distribution transformation proposed by Prade and Dubois as an example. This transformation will be used in some applications in this contribution, namely in the competition-based possibility and necessity measures aggregation, and in missing data estimation.

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: , , where , for any possibility or probability measure defined on  (in this case we say that  dominates *P*). Thus, transforming a probability measure into a possibility measure can be materialized by choosing a possibility distribution in  (the set of all the possible measures that dominate *P*). Dubois et al have proposed to add the following constraints in order to ensure the preservation of the distribution form: , where , and , for all . To reduce the imperfection of an information element, we must choose the distribution the most specific (in the fuzzy set theory we say that the possibility distribution  is more specific than  if , ). Dubois and Prade show that the solution to this problem exists and it is unique. The solution can be defined as the following:

Supposing that  , it is possible to define a strictly ordered relation on  such that: . Let  be a permutation of the indices  associated with the strict order: , or in another way: . The permutation  is a bijection and the inversed transformation  gives the rank of each  in the list of the probabilities reordered as an increasing sequence. Accordingly Dubois-Prade transformation can be given as:  .

For instance, let , then , , , and . Accordingly, , , , and .

Consequently:



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  has to be taken into account. For this purpose, this partially order is represented by a set of its linear extensions Λ*(P)={}.* At each possible linear extension from Λ*(P)*, there is a permutation  from the set  that corresponds to  in such a way that : . In this case, the distribution the most specific and compatible with  can be obtained by taking the maximum of all the possible permutations as: , 

For instance, let , there are two possible permutations in this case: , , , and ; and , , , and . By applying the transformation, we find the following:  ,

, .

Notice that  implies that  (this condition is imposed by the preservation of the strict order).

* **Possibilistic Similarity Modeling**

Suppose that  is a dataset that consists of *“n”* objects, and each object  is defined or characterized by a descriptor vector , where *“S”* is the number of the attributes (the features, the characteristics, the descriptors, etc.) that describe this object and   is the value of the *ith*attribute in . We aim to measure the similarity between any two objects  and  defined via the vectors  and  () in the general case where  or ,  can take qualitative, quantitative, or ordinal values which can be in its turn imprecise, ambiguous, and/or uncertain. We must also take account of the missing values (where  is not assigned for some *i*s) or even of the attributes that exist in some objects and don’t exist in the others. We will see in the following that the possibility theory introduced before can provides us with robust tools to model the similarity by two monotone measures, namely the possibility and the necessity degrees of similarity, taking into account all the aforementioned constraints in a very simple, human-reasoning-like, and intuitive way by getting use of the various robust reasoning techniques developed under the fuzzy set theory briefly illustrated in the next section.

* **Linguistic Variables & Fuzzy Propositions**

A *linguistic variable* is a 3-tuple information element , where *V* is the name of the variable defined on the universe  and the set  of the basic fuzzy characterization of *V*.

In order to assure an efficient reasoning from the values of *V*, we must construct more characterizations of *V*. This can be done by means of the *linguistic modifiers* (sometimes called *linguistic hedges*). A linguistic modifier is an operator “*mod”* that generates a new characterization *mod(A)* from any characterization *A* of *V* in such a way that  for a mathematical transformation  associated with *“mod”*.

For a set *M* of modifiers, *M(TV)* denotes the set of the fuzzy characterizations deduced from *TV*. In the following we present some well-known hedges with their mathematical definition:

* Very: .
* More or less: 
* Approximately:  for .
* About: , for a parameter  in [1/2,1].
* Rather: , with .

Supposing that *L* is a set of linguistic variables and *M* is a set of linguistic modifiers. For a linguistic variable  of *L*, “*V is A*” defined by means of a normalized fuzzy set *A* of  in *TV* or in *M(TV)*  is called an elementary or atomic fuzzy proposition. The compound fuzzy proposition is obtained by combining several atomic fuzzy propositions like “*V is A*” and “*W is B*”, etc. The simplest compound fuzzy proposition is a conjunction of elementary fuzzy propositions “*V is A and W is B*” for two variables V and W respectively defined on the universes and  (like for instance, the glycaemia level is abnormal and the cholesterol level is high”). It is associated with the Cartesian product  of the fuzzy sets of and *,* characterizing the pair on Its truth value is defined by  or more generally by  for a t-norm *T,* in any  of . Such a fuzzy proposition is very common in rules of knowledge-based systems and in fuzzy control. Similarly, we can combine elementary propositions by a disjunction of the form "*V* is *A* or *W* is *B*". The truth value of the fuzzy proposition is defined by  or more generally by  for a t-conorm , in any  of .

We can estimate the veracity of a fuzzy proposition “*V is A*” defined via , given a referential fuzzy proposition *“V is B”* defined via  using the possibility measure , and the necessity measure , defined as:





These two equations constitute the backbone of the possibilistic similarity.

* **Similarity Personalization & Tolerance Functions**

In data mining (*DM*) and knowledge discovery in databases (*KDD*), it is very useful to give the expert or the user the capability to represent his or her viewpoint by assessing to which extent he or she considers that two values of the same attribute (denoted as  and , , ) are similar. In order to model such similarity, Dubois and Prade have proposed an approach for the precise quantitative attributes that takes account of the user’s point of view by associating each attribute with a membership function modeling its similarity according to the user. This function can represent , , etc. defined on the domain . For example, suppose that we intend to measure the similarity between the record number *“1”* and each record *“2”*, *“3”* and *“4”* depicted in figure 17. Let us also suppose that each record contains two quantitative attributes and each attribute is associated with a membership function describing each similarity: the first attribute’s similarity is modeled by the function depicted in figure 17-a. According to this function, we consider that the two values of the first attribute are completely similar () if the difference between these values doesn’t exceed “2”. Beyond this value (“*2*”) the similarity attenuates gradually until “*4*”. When the difference exceeds “4”, the similarity becomes null (). The second attribute’s similarity is modeled by the function depicted in figure 17-b. According to this function, we consider that the two values of the second attribute are completely similar () if they are identical (their divisions equals to “*1*”), then the similarity gradually decreases and diminishes when. Table 2 contains the values of inter-attribute similarities, and figure 17-c depicts the inter-object (inter-record) similarities. Note that record “*3*” and record “*4*” are equivalently similar to record “*1*” since the differences between the values of the first attribute is still less than “*2*” and consequently they belong to the tolerance interval where we consider that the similarity is always equal to “*1*”. In this example, we considered that inter-object similarities are equal to the average of inter-attribute similarities.

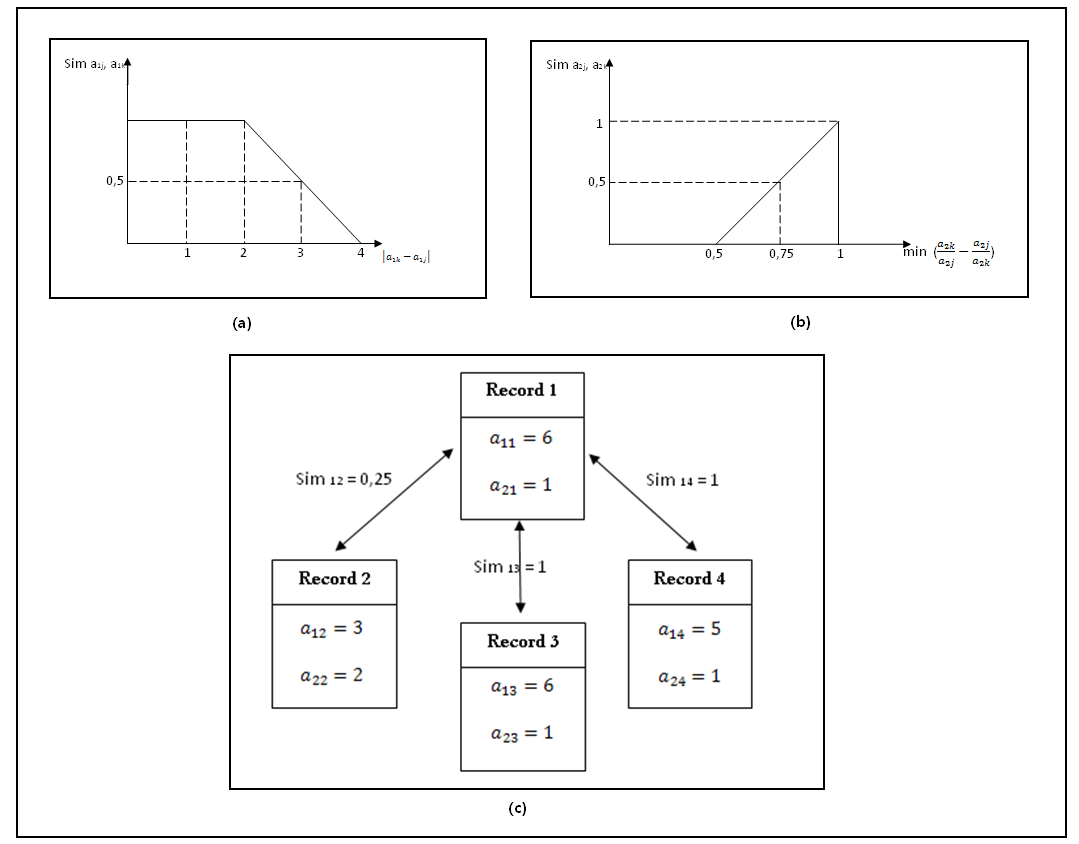


Figure 17. Illustrative example of similarity personalization.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Record “*2*” | Record “*3*” | Record “*4*” |
| Record “*1*” |  |  |  |

Table 2. inter-variable similarities.

Similarly, in our approach we’ll consider that each attribute is associated with a “*tolerance function*” defined by the expert or the user 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 a tolerance function is the function that we call “close to”. Such a function can be defined by the following formula:

 if 

 Otherwise

Where  is a variable that influences the slope of the function and consequently the notion of “close to”. The value of  depends on the nature of the attribute and on the user himself. For example, concerning the attribute “age”, for an expert the value of  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 the previous equation, whereas the value of  might be “10” for another expert more tolerant. The value of  might depends also on the domain of definition of the attribute. For example, for an attribute whose definition domain is , the value of  might be 1000, whereas for another attribute whose values belong to the interval , the value of  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 via tables to reflect their points of view about the similarities between the attributes.

* **Possibilistic Similarity Measuring**

Similarly to the example illustrated before we can measure the similarity between the descriptor vectors  and  by estimating the inter-attribute similarities and then by aggregating their values using the average or the possibilistic aggregation approach that we will propose at the end of this section. To do so, for each attribute, we consider that its associated tolerance function  is the fuzzy proposition that must be matched with the compound referential fuzzy proposition given as “the informational content of the attribute  in  is defined via the possibility distribution  and the informational content of the attribute  in  is defined via the possibility distribution ” that can be mathematically defined using the t-norm or the minimum as:



In this case we can calculate inter-attribute possibility and necessity degrees of similarity between the two fuzzy propositions using the following equations:





where .

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  but we are entirely uncertain . Whereas, if the attribute is defined in only object and completely undefined in the other, we consider that  (this takes place when the objects come from different sources, experts, hospitals, etc.).

Unlike many fuzzy-set-theoretic techniques and approaches that aggregate the membership degrees or the possibility, the necessity, or the belief masses using the minimum or the maximum, we propose to use the average (to provide a quantitative value of the similarity) or the possibilistic aggregation approach (to provide a qualitative description of the similarity) because they are more appropriate in the context of the similarity to avoid the domination of only one attribute.

Concerning the possibilistic approach to aggregate the possibility and the necessity degree, let us suppose that  is the set of all the possible values of local possibility (necessity) degrees of similarity and  is its cardinality.  is an ordered set in which each element represents the frequency of the corresponding element of which is calculated as follows:

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

.

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

If  is equal to  then 

Now that we built the set , we create the ordered normalized frequency set  of , where is calculated as 

Then we apply Dubois-Prade transformation in order to construct the possibility distribution set  . The definition domain of the possibility or the necessity degree denoted as  is divided into *C* fuzzy regions whose membership functions are chosen by the user. For example,  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  (or  when dealing with the necessity degrees) is the *k-th* fuzzy region and that  (or  for the necessity) is its membership function (). We calculate the membership degrees of each element (value) of the set *V* to each fuzzy region (or to ), denoted as( or as  for the necessity). For each fuzzy region we calculate the possibility or the necessity membership (the possibility that the similarity between  and  belongs to the considered region) by means of the following four equations:



a). For the necessity membership degree ():

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





b). For the possibility membership degree ()

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





The first equation represents the necessity degree that the two records belong to the region  given that  represents all the possible values of their local possibility degrees, whereas the second equation represents the necessity degree that the two records belong to the region  given that  represents all the possible values of their local necessity degrees. The third equation represents the possibility degree that the two records belong to the region  given that  represents all the possible values of their local possibility degrees, whereas the fourth represents the possibility degree that the two records belong to the region  given that  represents all the possible values of their local necessity degrees. We must note here 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  and  belongs to the fuzzy region whose possibility (necessity) membership degree is the maximum. In other words, in this approach, the decision concerning the similarity can be done in four different ways according to , , , or to . In the following, we give two simple concrete examples. The first one concerns the similar records while the second deals with dissimilar ones.

* **Illustrative Example 1**

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 3 as follows:







As we can see, and . Consequently, we can conclude that the two records are quite similar (as expected). In this example: , , , .

* **Illustrative Example 2**

Suppose 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 4 as follows:







As we can see,  and . Consequently, we can conclude that the two records are dissimilar (as expected). In this example: , , , .

Table 3. An example of two very similar records (Example1).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *V* (all possible) | 0.70 | 0.75 | 0.80 | 0.90 |
|  | 0.20 | 0.30 | 0.40 | 0.10 |
| (possibility distribution) | 0.30 | 0.60 | 1 | 0.10 |
|  | 0.70 | 0.40 | 0 | 0.90 |
|  | 0 | 0 | 0 | 0 |
|  | 0.30 | 0.25 | 0.20 | 0.10 |
|  | 0.70 | 0.75 | 0.80 | 0.90 |
|  | 0.70 | 0.40 | 0 | 0.90 |
|  | 0.70 | 0.40 | 0.20 | 0.90 |
|  | 0.70 | 0.75 | 0.80 | 0.90 |

Table 4. An example of two very dissimilar records (Example 2).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *V* | 0.10 | 0.15 | 0.30 | 0.40 |
|  | 0.20 | 0.50 | 0.20 | 0.10 |
|  | 0.50 | 1 | 0.50 | 0.10 |
|  | 0.50 | 0 | 0.50 | 0.90 |
|  | 0.90 | 0.85 | 0.70 | 0.60 |
|  | 0.10 | 0.15 | 0.30 | 0.40 |
|  | 0 | 0 | 0 | 0 |
|  | 0.90 | 0.85 | 0.70 | 0.90 |
|  | 0.50 | 0.15 | 0.50 | 0.90 |
|  | 0.50 | 0 | 0.50 | 0.90 |

* **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. presented briefly as follows:

Supposing that 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 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  is the frequency of appearance of the modality  in the set of cases belonging to the class “*A*”, and  is the set of the modalities that describe the case .  is the membership degree to class “*A*” calculated as:



The similarity can be calculated as:



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).

* **Experimentation**

With each attribute of the objects of the medical dataset illustrated in the appendix we associated a tolerance function as a first fuzzy proposition in such a way that for the numerical values, “close to” functions with  were assigned, and their values were considered as imprecise information modeled by symmetric fuzzy triangular numbers (this resembles to adding white noise to a signal in signal processing). These “close to” functions were also associated with the ordinal attributes in order to personalize and to control the similarity between a category or a property and its neighboring properties. Finally, with the nominal and binary attributes we associated “true and false” functions. Then, the inter-attribute possibility and necessity degrees of similarity were calculated and aggregated using the steps introduced in the previous section. Afterwards, we get the inter-object possibility and necessity similarity matrices that respectively contain the upper (optimistic) and the lower (pessimistic) limits of the similarity between the patient records. Later, we will try to graphically and geometrically visualize the content of the similarity necessity degree matrix and to compare the conclusions that we can draw from the representational models with the ground truth provided by the expert, and then this matrix will be exploited in many application in data mining and knowledge discovery (as pattern recognition, clustering, missing data estimation, self-organizing maps, etc.). We will show that in all the cases, this matrix provides us with results and conclusions identical to the ground truth assigned by the doctor.

* **Similarity Representational Models**

The meaning "to visualize" does not correspond to the action to see which means “to perceive by the meaning of the view” but to that of «to make perceptive to the view what is not naturally visible". In fact, Visualization is the process of transforming invisible abstract data, information, and knowledge into a visible display in the form of geometric or graphical representations in order to support tasks such as data analysis, information exploration, trend prediction, pattern detection, rhythm discovery and so on. Actually, the representational models give observed events a meaningful interpretation and allow future or unseen events to be anticipated through the process of generalization. In order to represent the possibilistic similarity proposed, we have chosen well-developed mathematical representational models like the linear and the circular unidimensional scaling (LUS and CUS) and the multidimensional scaling models (MDS) for similarity spatial visualization, then we have chosen the additive and the ultrametric trees for the graph-based visualization. We must indicate here that it is not the intent here to present formal demonstrations of the structural representations such as the mathematical algorithms or the convergence of these approaches. All of this is generally available in the literature. The primary interest here is to present and to demonstrate that these models can be applied as they are to the possibilistic similarity matrix as well as to the conventional proximity matrices, and the results remain coherent, representative, and as expected. The computational approach implanted for obtaining all the sundry models are based on what is called the iterative projection method for solving linear inequality constrained least-squares tasks. All the presentations of concern can be characterized by explicit linear inequalities; thus, once the latter constraints are known, the actual representing structure can be obtained by using the iterative projection. These methods are applied for the first time to this type of similarity matrix.

* **Spatial Visualization**

Spatial visualization (sometimes called geometric representation) is essentially based on the scaling techniques. Scaling consists of measuring and comparing objects in some meaningful way based on mapping whose goal is to locate or to map these objects based on their similarity onto a smallest space that will effectively retain the basic information about the data. This projection may reveal the underlying structure or the unique relationships among the items. One of the most known linear representations used in our daily life is Fahrenheit scale. Gabriel D. Fahrenheit had created a relative scale for assigning temperatures to mercury heights by dividing the distance between the boiling and the freezing marks of water into 180 equal parts or units. In our experiments, we will try to assign the estimated values of the possibilistic similarity to the geometric distances between the points represented in an n-dimensional space.

**\*\* Linear and Circular Unidimensional Scaling (LUS and CUS)**

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 object locations 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.

**\*\* Multidimensional Scaling (MDS)**

MDS model represents  objects as points in an  coordinate space denoted as  so that the distances between points  approximate a given object distance matrix . MDS algorithms generally consist of the iterative minimization of a stress function measuring the discrepancies between the given distance matrix (in our case the similarity necessity matrix) and the reconstructed distances in the configuration space. The various methods available differ by the choice of the stress function and the optimization algorithm used. Herein, we apply gradient descent algorithm to minimize the sum squared error that represents the stress function given as:



Where  is measured according to one of Minkowskian *p*-metric:



The learning rule in this case is given according to [] as:

Where is the signum function, and  is the learning rate parameter.

* **Graphical Visualization**

Graphs and their visualizations are essential tools in data exploration and understanding, particularly for those applications that need to manage, process, and analyse huge quantities of data. The benefits of graph drawing techniques for data mining purposes include analysis through visual exploration, discovery of patterns and correlations, abstraction and summarization. Additive and ultrametric trees (briefly presented below) are two efficient well-known types that get use of graph theory to visualize the structure of patterns in a specific dataset.

**\*\* Additive Trees**

This algorithm is very similar to multidimensional scaling algorithm and is founded on a search 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. Herein, we use once again the population-based incremental learning algorithm (PBIL) to minimize the stress function given as: .

**\*\* Ultrametric Trees**

An ultrametric tree is a special kind of a rooted additive tree where the terminal nodes are all equally distant from the root. This type of trees is widely used in hierarchical clustering and is referred to as the dendrogram.

* **Visualization Experimental Results**

In the following we will apply the aforementioned conventional models of representations to the similarity necessity matrix of the digestive image database calculated before in order to display its abstract contents. As we will see, these models are capable to obviously show the classes of the pathologies of the images and the relationships that exist between the object of the same or of the different pathologies. In order to clarify our visual representation, the each representational model will be applied twice. The first time, to a small dataset , where  (remark that we always make reference the notations described in appendix *I*), and the second time to .

Table 5. LUS applied to .

|  |  |  |
| --- | --- | --- |
| Objects | coordinates | Pathology |
|  | -0.6810 |  |
|  | -0.5895 |
|  | -0.4640 |
|  | -0.4035 |
|  | -0.2676 |  |
|  | -0.1796 |
|  | -0.0453 |  |
|  | 0.0627 |
|  | 0.1761 |
|  | 0.2672 |
|  | 0.3603 |
|  | 0.4779 |
|  | 0.5813 |
|  | 0.7050 |

Table 6. LUS applied to .

|  |  |  |
| --- | --- | --- |
| Objects | coordinates | Pathology |
|  | -0.7285 |  |
|  | -0.6255 |
|  | -0.5389 |
|  | -0.4511 |
|  | -0.3758 |
|  | -0.3041 |
|  | -0.2156 |
|  | -0.1259 |
|  | 0.0004 |  |
|  | 0.0448 |
|  | 0.1228 |
|  | 0.2012 |
|  | 0.2999 |  |
|  | 0.3733 |
|  | 0.4541 |  |
|  | 0.5382 |
|  | 0.6185 |
|  | 0.7123 |  |

Using the algorithm of LUS applied to the possibilistic similarity matrix in order to represent it along a linear continuum, and the algorithm of CUS applied to represent this matrix in the circular continuum, we get the results presented in tables 5 and 6 concerning the linear unidimensional scaling, and in figures 18 and table 7 for the circular unidimensional model. 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 in this context. From the constructed matrix or/and from the obtained categories and coordinates in the tables 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.

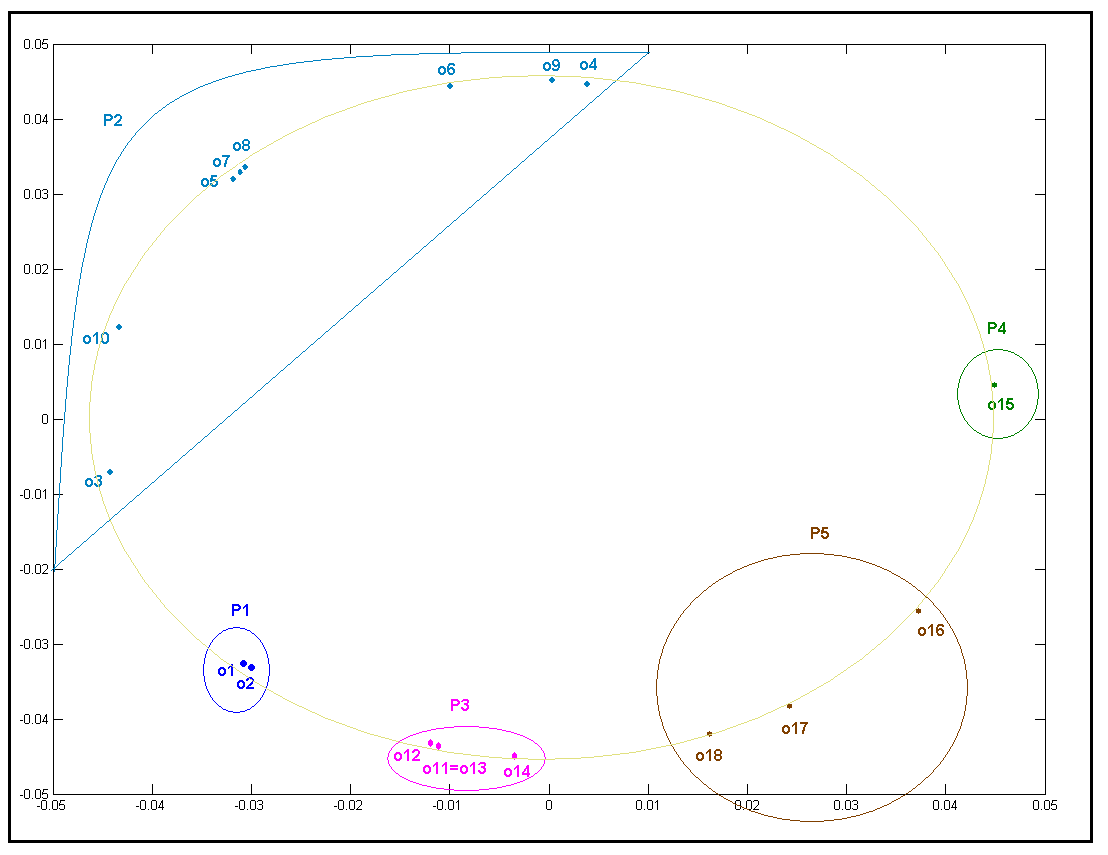


Figure 18. CUS applied to .

Table 7. CUS applied to  (the coordinates).

|  |  |  |
| --- | --- | --- |
| Objects | coordinates | |
|  | -0.0311 | -0.0335 |
|  | -0.0301 | -0.0345 |
|  | -0.0444 | -0.0073 |
|  | 0.0037 | 0.0449 |
|  | -0.0309 | 0.0328 |
|  | -0.0101 | 0.0439 |
|  | -0.0308 | 0.0330 |
|  | -0.0306 | 0.0350 |
|  | 0 | 0.0450 |
|  | -0.0433 | 0.0124 |
|  | -0.0111 | -0.0436 |
|  | -0.0114 | -0.0433 |
|  | -0.0111 | -0.0436 |
|  | -0.0037 | -0.0449 |
|  | 0.0448 | 0.0041 |
|  | 0.0371 | -0.0256 |
|  | 0.0243 | -0.0379 |
|  | 0.0159 | -0.0421 |

Let us now apply the 3-dimensional multidimensional scaling (*3D MDS*) algorithm to the possibilistic proximity matrix (see figures 19 and 20). The same remarks can be deduced concerning the similarity of objects (remark for example that the objects *O11* and *O13* are the object the most similar in this base). Note that though we have presented the 33D objects with just 3D space, the representation is still coherent and gives the expected results, and it becomes accessible to the visual inspection and exploration.

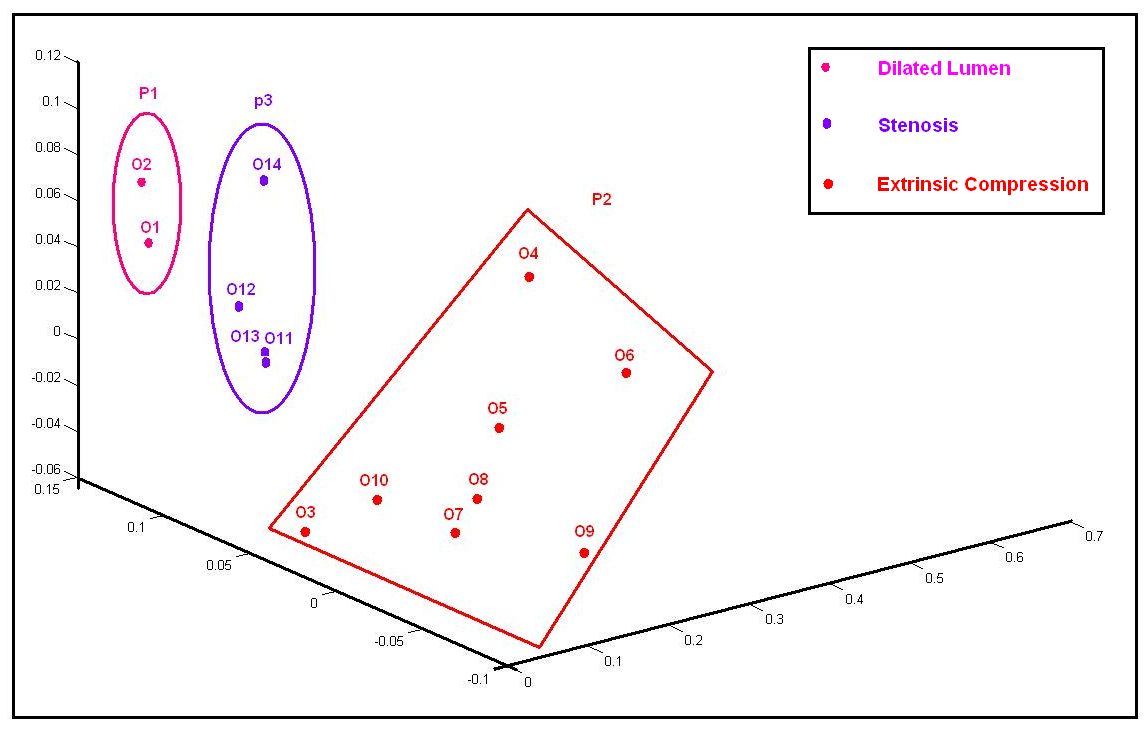


Figure 19. MDS applied to .

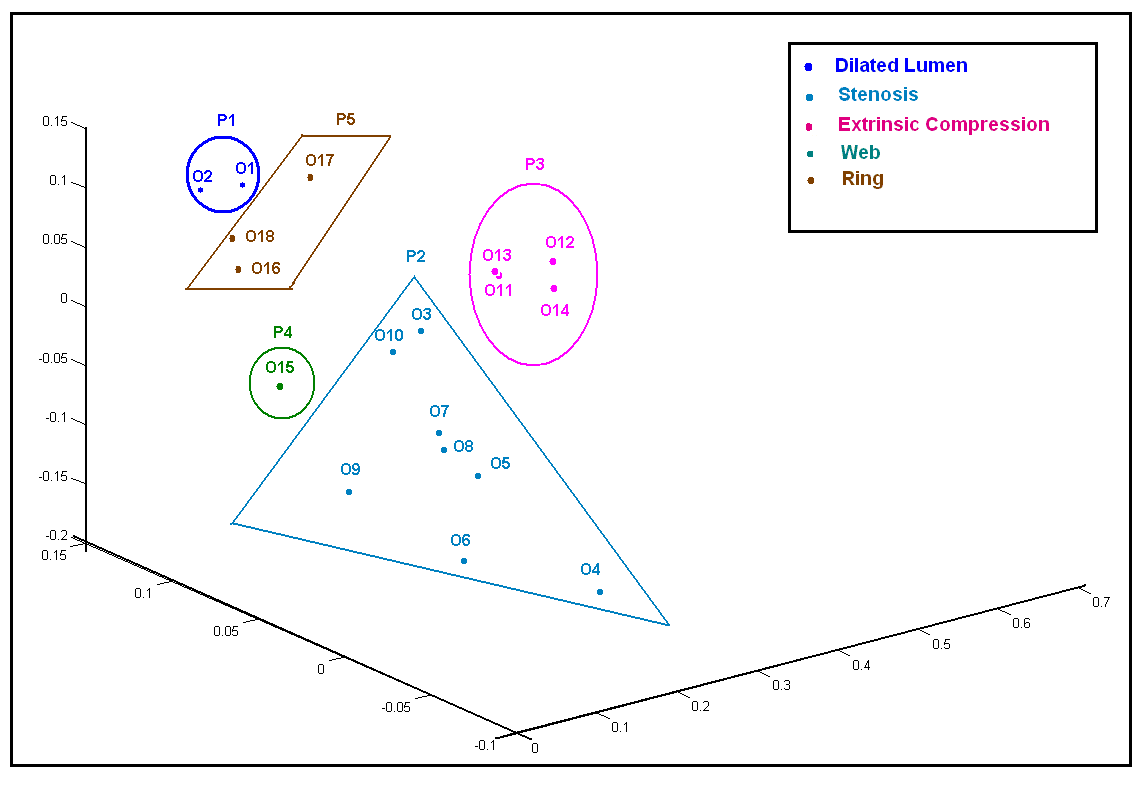


Figure 20. MDS applied to .

The possibilistic proximity matrix can be also represented as additive trees and ultrametric trees using the graph theory techniques, and similar results and conclusion could be obtained. In these methods the objects belonging to the same pathology are generally attached to the same internal node. Figures 21, 22, 23, and 24 gives examples of applying the additive tree model to the both databases using different numbers of the internal nodes that represent the cluster of objects of the same class. Remark in these figures that if the number of the internal nodes is less than the number of classes (the number of the pathologies), then objects of different pathologies may gather together and conduct the same internal node. Only the pathology the most distinct can be firstly separated. By increasing the number of the internal nodes, the other distinct pathologies will be separated. However, when the number of the internal nodes exceeds the number of the classes, subclasses of the main class can be born (for example from the class hernia, the subclasses hernia type 1, type 2, and type 3 can be generated). This characteristic of the additive trees is very interesting to enable the analyst to have an idea about the pathologies the most distinct of his database, or the pathologies the most similar which are always attached to the same node. Notice here that we are not only capable to estimate the similarity between the pathologies, but we can also estimate the similarity between the object via the length of the path that connects these two objects.

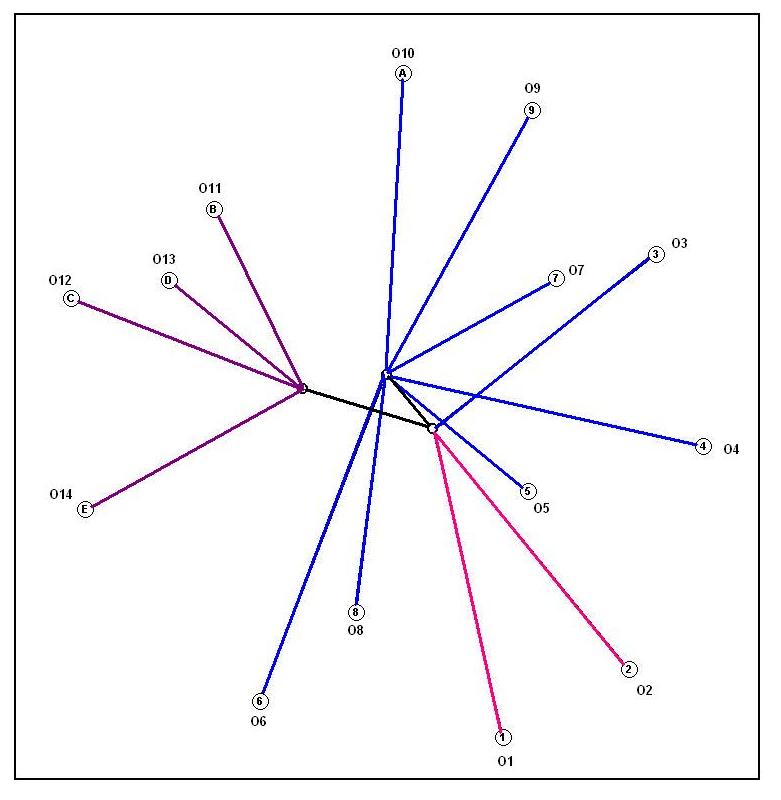


Figure 21. a 3-internal-node additive tree representation of 

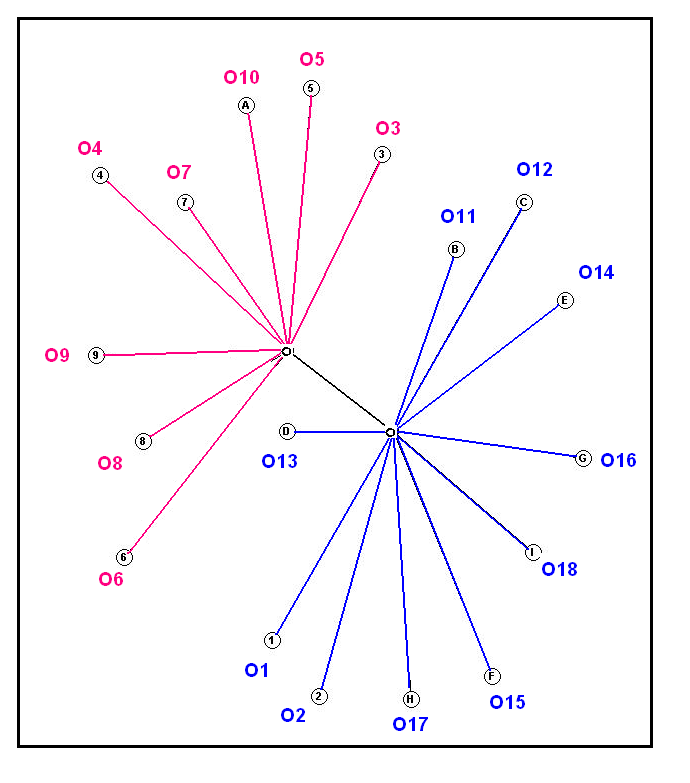


Figure 22. a 2-internal-node additive tree representation of 

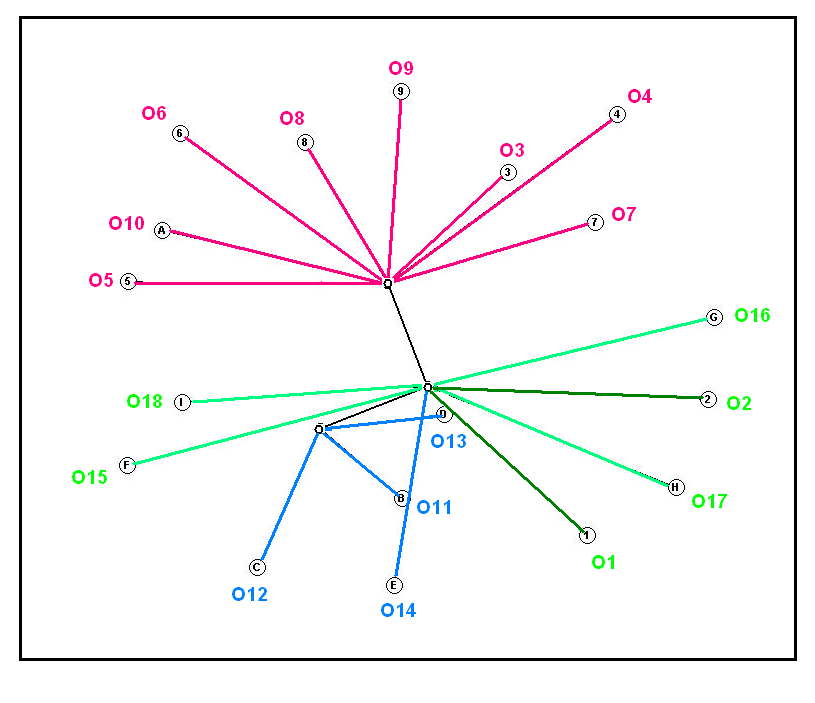


Figure 23. a 3-internal-node additive tree representation of .

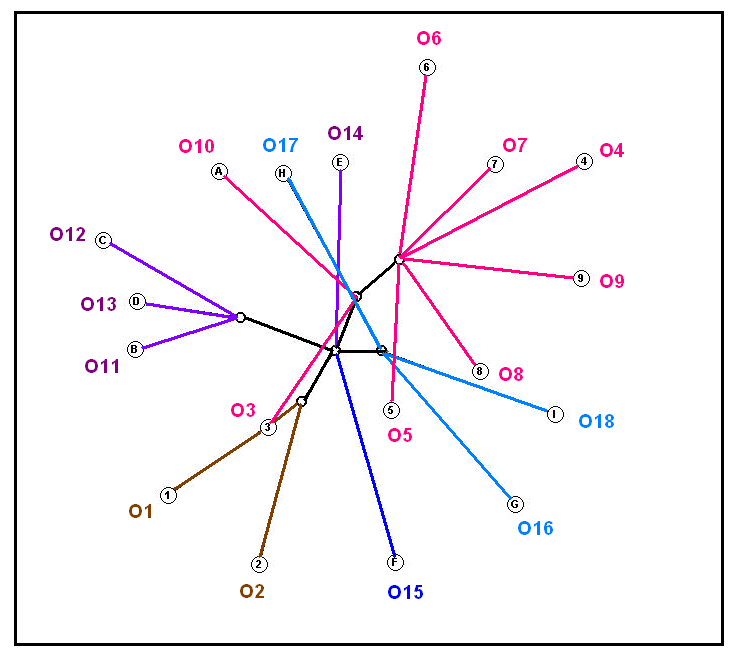


Figure 24. A 6-internal-node additive tree representation of .

Actually, in addition to the visualization models presented above, similarity measures can be represented using the clustering techniques and this model is called structural visualization. This type of visualization will be illustrated later with examples. As we might remark here, the heavy reliance on graphics to represent the data gives the analysts unparalleled power to explore the dataset by enticing the data that compose it to reveal its structure secrets in order to gain some new, often unsuspected, insights in these data. This insight implies detecting and uncovering the potential structure in the data in order to: 1- test underlying assumptions (remark that we can easily determine the pathology class of each object (image) by just having a glance at the graphics, 2- extract important variables (like extracting the key attributes of each pathology using the association rules applied to the local degrees of possibility and necessity), 3- detect the outlier and the anomalies in the data.

In this way, we can sum up that we established an interface “a communication channel” between data held in artificial systems, and the humans who must deal with this information, by using conventional visualization methods and techniques applied to possibilistic similarity matrix. These models do not only help the analyst to comprehend and to remember his or her information, but also they help to quickly locate the tendencies in data and as well as the anomalies. Each of theses visual representations give a different image of information from different angles of interpretation and can mean a thousand of words.

1. **Possibilistic Data Mining Applications**

In the following, we propose some mathematical frameworks unified under the unified umbrella of possibility theory in medical data mining domain to deal with imperfection in an efficient way.

* **Imperfect Pattern Recognition**

In spite of the large number of studies on pattern learning information processing, there are until now some crucial problems related to information imperfection, heterogeneity, conflict, scarcely, and scalability, still unsolved within a unified framework. Information imperfection can take place either when labelling the patterns in the training set via an automatic system, or when assigning imprecise, ambiguous, uncertain, or even missing values to pattern features. Unlike the first type of imperfection that has been considered in the recent studies, feature imperfection is still disregarded in the literatures. This issue can lead to a high difficulty in measuring the similarity between the patterns; which is a pivotal task in recognition; due to the limitations or the constraints of the conventional measures, especially when taking account of the different data types (quantitative, qualitative, ordinal, etc.) in the same context with the imperfection. In addition to imperfection, information conflict and scarcity have to be considered. The first issue could take place when a pattern is close to several training examples from different classes, whereas the second arises when a pattern is far away from any pattern in the training set. In this case, the pattern most probably belongs to unknown class for which no information has been gathered in the training set, and should be rejected. Finally, as pattern recognition could be used as a pre-processing step proceeding other data mining tasks like rule extraction, anomaly detection, etc., the output information of the constructed recognition system have to be flexibly transformable to other forms suitable for the next steps.

Next, we introduce our approach that takes account of all the aforementioned design points and conditions by getting use of some monotone measures like possibility, necessity, belief, and plausibility degrees, presented in the united framework of the fuzzy measure theory. Then, a concrete example - aiming to clarify and to numerically explain the steps of the proposed system- is presented. This approach will be applied to a real gastroenterology endoscopic image database described in appendix I.

* **Possibilistic Evidential Pattern Recognition System**

Let us suppose that we want to classify a set of observed patterns  depending on a training pattern set  into the class set , given that *A* represents a multivariate feature vector of the considered pattern, whose variables could have imperfect (imprecise, ambiguous, uncertain, or even missing) values, and could be heterogeneous (quantitative, qualitative, ordinal, etc.). We will also suppose that the class membership of the training pattern is assessed by an expert or by another automatic system via a possibility distribution. Specifically, each example  for  is associated with a possibilistic label modeled as: , in such a way that  and  (see figure 25). The proposed system consists of two main phases.

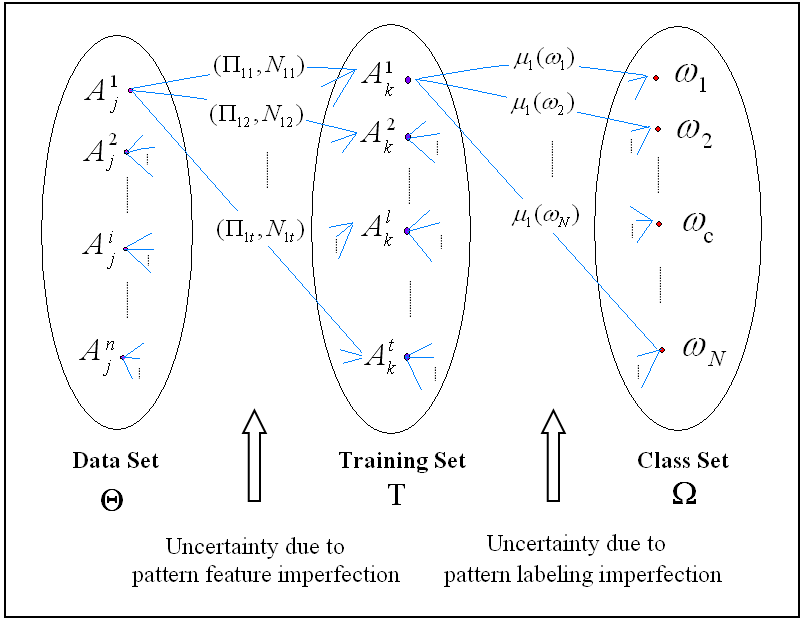


Figure 25. Imperfect pattern recognition system.

The first one concerns measuring the similarity between two patterns having imperfect heterogeneous features based on the fuzzy measures of possibility theory explained above. The second step deals with the imperfect knowledge of the training set getting use of the monotone measures of Dempster-Shafer theory processed as follows:

After calculating the degrees of similarity between  and , denoted as and ,  and , we calculate the basic belief assignments of the observed patterns as the following algorithm:

For each , 

For each , 

 , 

(represents the possibility-based evidence, and  represents the necessity-based evidence, provided by the example  concerning the membership of the observed pattern  to all the classes of )

As we might remark, each training pattern plays the role of an item of evidence (witness) influenced by similarity possibility and necessity degrees, on the ground that the observed pattern tends to belong to the class of the training pattern with the same membership degree if it is sufficiently similar to this training example.

Using any fusion rule from the literature (Dempster-Shafer, Yager, Dubois-Prade, etc.), we can combine the possibility-measure-based belief and the necessity-measure-based one as:

 ( represents our belief in the class membership degree of )

where  and .

Now that the basic belief assignments are calculated, the decision concerning the pattern class can be carried out using the maximum belief (credibility) value of the classes of , or the maximum plausibility value, or as a compromise between these values, we can consider the pignistic probability value.

* **Numeric Example:**

Figure 26 gives an example of classification an observed pattern (record for instance) according to the testimonies given by two training evidence items, combined according to Dempster-Shafer fusion rule. We suppose in this example that each pattern (record) is characterized by two attributes. The first attribute takes imprecise values, while the second one is defined by a probability distribution resulting from a certain rule based on the first attribute or based on the output of another process or system. As we can notice in this very simple case, calculating the similarity using the conventional proposed measures is complex, difficult, and even impossible. While, it is straightforward, simple, and easy thanks to the possibilistic approach that we proposed above. Actually, for the next attribute, we must transform the probability distribution to possibility distribution using any transformation rule proposed in the literature. In this example we used Prade-Dubois transformation. According to its similarity with the observed record and to its class membership degree, each training pattern,  gives an optimistic evidence  concerning the class membership degree of the observed pattern based on the similarity possibility degree and a pessimistic evidence  based on the similarity necessity degree. Afterwards, these evidences are combined using Dempster-Shafer rule for example to calculate the final belief masses. Now that all the basic belief masses are calculated, we can take the evidentional decision by considering either the singleton maximum belief (in the optimistic case), or the single maximum plausibility (in the pessimistic case), or even by considering the maximum pignistic probability. According to the values shown in Figure 26, we can decide that the observed pattern belongs to  in all the aforementioned criteria.

* **Imperfectly-Described Image Recognition:**

Our approach has been applied to a large gastroenterology endoscopic image database, in which the lesions presented in each image (pattern) is described by an expert via up to 33 attributes, with 206 different modalities (we have at least 8 imprecise attributes, and up to 6 missing values in each pattern). This base is well detailed in the appendix. Each training pattern is assigned to one class (has one label concerning the lesion’s type). Accordingly, the output belief masses of the proposed system are directly related to pattern possibilistic similarity values. Therefore, in order to materialize our results, we can visualize this possibilistic similarity (that is directly related to basic belief masses) by two representational models: a *spatial model* based on the circular unidimensional scaling for example that aims to represents the objects along a circular continuum in such a way that the distances between them reflect their dissimilarity, or a *graphical model* using an additive tree for example where the shortest path between any two external nodes (patterns) represents their dissimilarity. As we have seen before, the patterns that belong to the same pathology gather together in such a way that the dissimilarity between any two objects of the same class is smaller than the dissimilarity between two objects belonging to different classes. Correspondingly, the similarity will be higher, and their corresponding basic belief will be larger, regardless of the objects chosen to constitute the training or the test set.

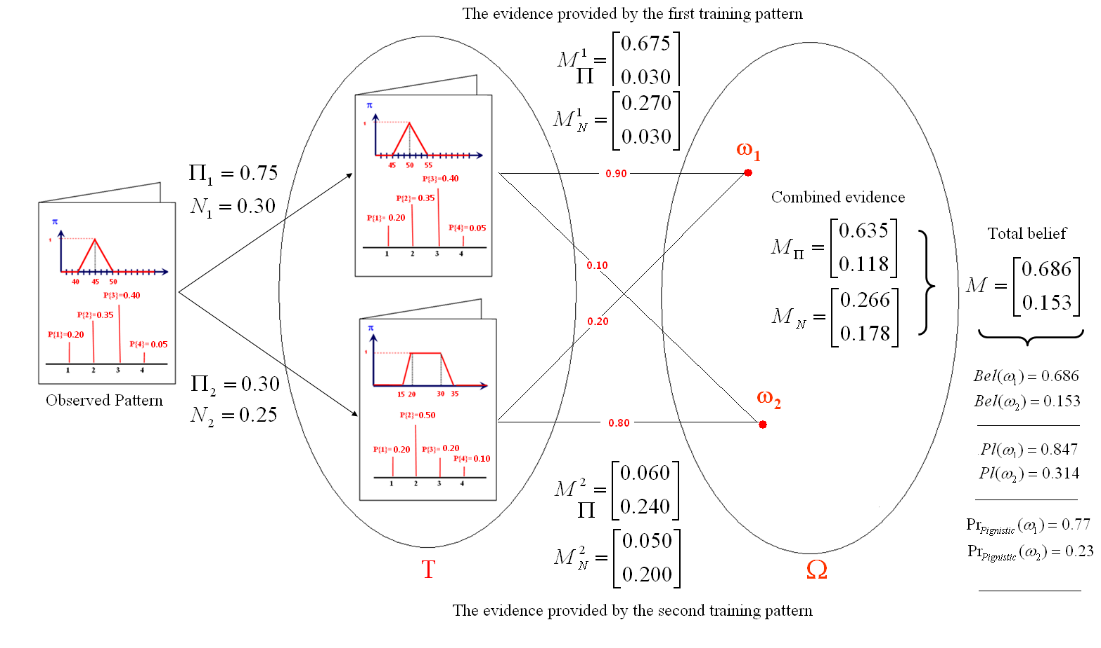


Figure 26. A simple concrete example of possibilistic evidential pattern recognition.

The proposed method takes account of information uncertainty in both pattern labeling and pattern description through the possibility theory measures and cope with the problems associated to information conflict and scarcity by putting the total evidence on  in the case of the inlier and outlier patterns, using Dempster-Shafer theory to combine the information resulting from evidences provided by all the examples of the training set weighted by possibilistic measures. Actually, the ability to convert any evidential partition to possibilistic partition (by calculating the plausibility measures of the concerned belief function) or to fuzzy partition (by applying the Pignistic probability to the singleton masses), or even to hard partition (by taking the maximum of the fuzzy partition), assures the scalability of the recognition system. Furthermore, the basic mathematical operators (max, min, sum, etc.) on which this method is essentially based on ensures a fast computation time, which is an important issue in data mining. Finally, thanks to its generality, this approach can be applied without modification to any other database.

1. **Possibilistic Kohonen Self-Organizing Maps**

Kohonen networks (known as self-organizing maps SOMs) are an effective mechanism in signal processing. They can convert a complex high-dimensional input signal into a simpler low-dimensional discrete map. Ritter has shown that SOMs represent a nonlinear generalization of principal components analysis (another dimension-reduction technique). Thus they are nicely appropriate for cluster analysis, image and sound processing, and many other applications. Actually, Kohonen networks are a special type of the neural networks based on the competitive learning which is based on similarity estimation. All the previous works and applications of the SOMs suppose generally that the value of each input is precise, certain, and given in order to calculate the similarity and to estimate the new weights of the network, while in reality, a remarkable amount of incomplete and imperfect values may be presented to the input of the artificial neural networks. For this reason we propose to apply the possibilistic similarity to overcome this problem. Next section presents briefly the SOMs, the algorithm, and the limitations. Then, we will propose an approach to overcome these constraints, illustrated via a concrete example.

* **Self-Organizing Maps**

Kohonen networks were introduced in 1982 by Finnish researcher Tuevo Kohonen as a special type of neural networks to reduce the dimensionality of the input signals. Like neural networks, SOMs are feedforward and completely connected. Feedforward networks do not allow looping or cycling. Completely connected means that every node in a given layer is connected to every node in the next layer, and unconnected to any node in the same layer. Each connection between nodes has a weight associated with it, which is assigned randomly to a value between zero and one at initialization. Adjusting these weights represents the key for the learning mechanism. Input variable values need to be normalized or standardized so that certain variables do not overwhelm others in the learning algorithm. Unlike most neural networks, SOMs have no hidden layer. Data from the input layer is passed along directly to the output layer. The output layer is represented in the form of a lattice whose shape is usually rectangular (see figure 27) or hexagon.

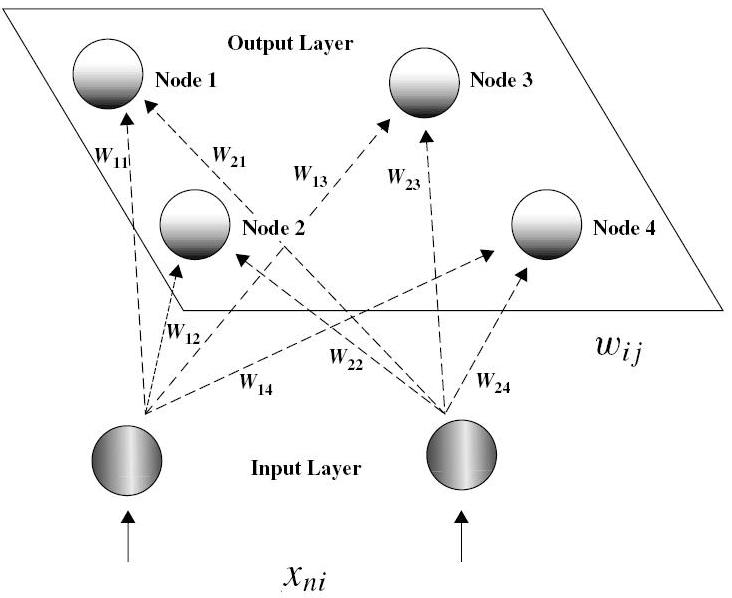


Figure 27. Self-organizing map topology.

For a given object (record, instance, stimuli, feature vector), a particular field value (attribute, variable, observation, feature, samples, examples) is forwarded from a particular input node to every node in the output layer. The values of all the fields, together with the weights assigned to each connection, determine the values of a scoring function (such as Euclidean distance) for each output node. The output node with the best outcome from the scoring function would then be designed as the winning node. This node becomes the center of neighborhood of excited neurons whose weights are adjusted so as to further improve the score function. In other words, these nodes will participate in the adaption (learning) process.

* **Kohonen Algorithm**

For each input vector from the data set, defined as:



where ,  represents an object in the data set like a patient record in a medical database for example and , , represents an attribute of  (like the hemoglobin or the age of the patient), do:

I- Competition:

For each output node , calculate the similarity (or the dissimilarity) between the two vectors  (the weight vector associated to this node) and  (the input vector) using the scoring function (the Euclidean distance given as  ) and find the wining node  that maximizes the Similarity (or minimizes the dissimilarity) over all the output nodes.

II- Cooperation:

Identify all output nodes  within the neighborhood of  defined by the neighborhood size . For these nodes, do the following for all input record fields:

1. Select the nodes in the neighborhood of the winning node that will participate in the learning phase. The weights of these nodes are adjusted so as to further improve the score function. In other words, these nodes will have an increased chance of winning the competition once again, for a similar set of field values.
2. Adjust the weights as:.
3. Adjust the learning rate and neighborhood size, as needed.
4. Stop when the termination criteria are met.

The majority of the previous works and studies that utilize Kohonen networks consider that the values of the attributes of the data set are crisp and observed (given), supposing that in the worst case the samples (attributes) can be cleaned and prepared in order to get a perfect training set. This optimistic point of view can not be applied to a great deal of the everyday databases for two main reasons. On the one hand, there are always some attributes that cannot be estimated exactly (precisely) because of the measuring instrument tolerance, or the expert uncertainty and doubt, so they can be given as a vague or as imprecise values modeled by possibility distributions. On the other hand, it is common to find missing values of the attributes in a data set. Deleting the fields or the attributes that contain such values might decrease the size of the learning base to a great deal plaguing the learning process. Estimating the values of the missing data before the learning process might be complicated, long, and uncertain. For these reasons, among others, we propose to replace the scoring function by the possibility and necessity degrees of the similarity in order to be able to deal with the imperfection of the data, and to present another equation for weight adjustment that takes into account all the possible values of a certain attribute according to its possibility and necessity degree.

The total possibility (necessity) degree of a certain node  is the average of all the local possibility (necessity) degrees connected to this node. Generally, the weights are given as a vector of precise values and the input attributes could take imprecise values modeled by a possibility attribution like  the new weight is calculated as:

****

* **Concrete Example**

Consider the following simple example. Suppose that we have a data set with two attributes, age and income, which have already been normalized (see table 8), and the knowledge about these attributes might be imprecise modeled by a possibility distribution. Suppose that we would like to use a 2 × 2 Kohonen map to represent hidden clusters in the data set. Therefore, we would have the topology shown in Figure 28. This type of data set can not be solved using the traditional scoring functions that suppose that the compared values have to be crisp. Besides, we can not manage to apply the weight adjustment equation with such imprecise attributes, given as possibility distributions.

Table 8. The dataset of the illustrative example (SOM).

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | is about 0.8 (figure 9-2-a) |  | Older person with high income |
| 2 |  |  | Older person with low income |
| 3 | is about 0.2 (figure 9-2-b) | is somehow high (figure 9-2-c) | Younger person with high income |
| 4 |  |  | Younger person with low income |
| 5 | is given by its possibility distribution  (figure 9-2-d) |  | Older person with high income |

|  |  |
| --- | --- |
| (a) | (b) |
| (c) | (d) |

Figure 28. Possibility distribution of the imprecise values.

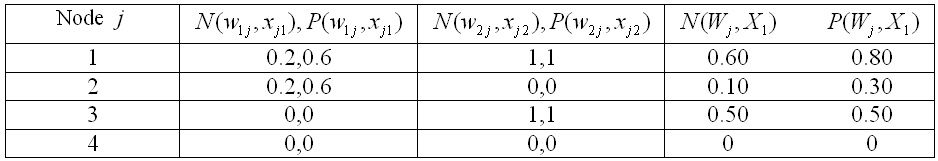
With such a small network, we set the neighborhood size to zero (), so that only the winning node will be awarded the opportunity to adjust its weight. Also, we set the learning rate  to 0.5. Finally, assume that the weights have been randomly initialized as follows:



For the first input vector, we perform the following competition, cooperation, and adaptation sequence:

A. Competition: We compute the necessity and the possibility degrees of similarity between this input vector and the weight vector for each of the four output nodes (see table 9):

Table 9. the necessity and the possibility degrees (the first vector).



The winning node for this first input record is therefore “node 1”, since it maximizes the similarity (modeled by the possibility and the necessity degrees) between the input vector for this record, and the weight vector, over all nodes. Node 1 won the competition for the first record because its weights are more similar to the field values for this record than the other nodes’ weights. For this reason, we may expect node 1 to exhibit an affinity for records of older persons with high-income. In other words, we may expect node 1 to represent a cluster of older, high-income persons.

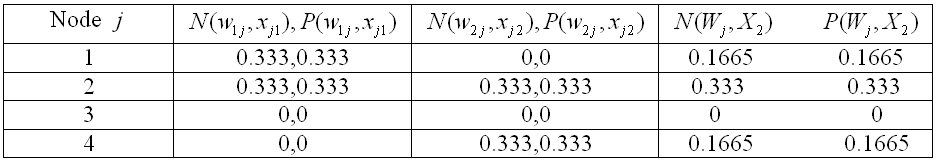
B. Cooperation: In this simple example we have set the neighborhood size R = 0 therefore, only the winning node, “node 1”, will be able to adjust its weights.

C. Adaptation: For the winning node, “node 1”, the weights are adjusted as follows:

, 

For the next input vector, see table 10

Table 10. The necessity and the possibility degrees (the second vector).



Node 2 won the competition because its weights (0.9, 0.2) are more similar to the field values for this record than the other nodes’ weights. As a result:

, .

For the third input vector, see table 11:

Table 11. the necessity and the possibility degrees (the third vector).

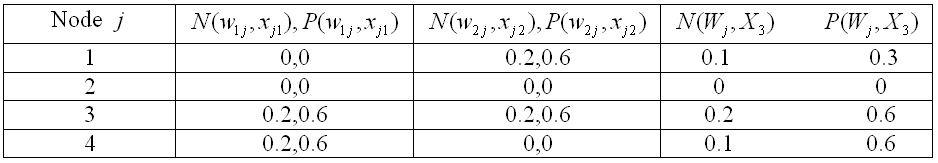


Figure 29 shows us the main steps of possibility and necessity degree calculation for the third training example.

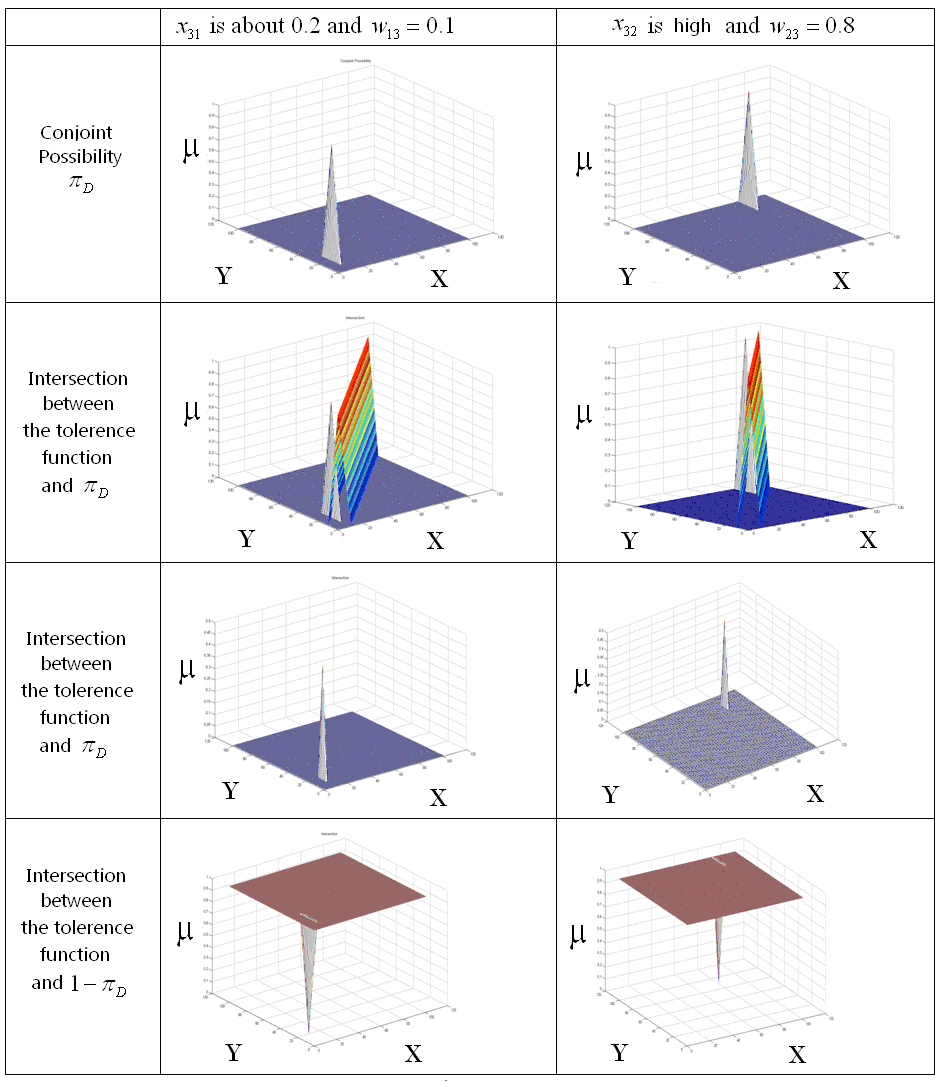
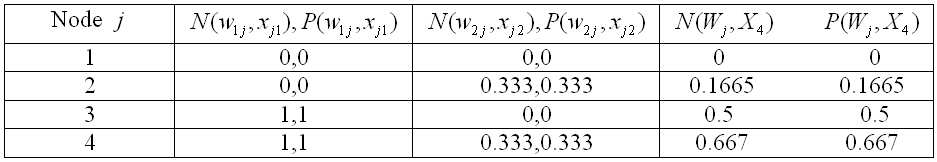


Figure 29. Local possibility and necessity calculation. X represents the first fuzzy proposition concerning the value of the attribute. Y represents the network weight. μ represents the possibility or the necessity degree.

Node 2 wins the competition, and as a result:, .

For the vector , see table 12.

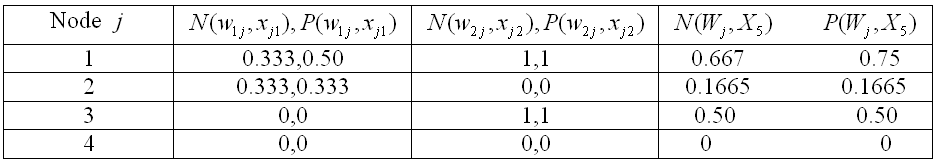
Table 12. The necessity and the possibility degrees (the fourth vector).



Node 3 wins the competition, and as a result:, 

Finally, for , see table 13:

Table 13. the necessity and the possibility degrees (the last vector)



Node 1 wins the competition, so given that  and  is given as possibility distribution  (figure 28-d), the new weight  is calculated as following:



Given that  and  the new weight  won’t change.

Kohonen networks are a type of the unsupervised artificial neural networks that are trained to produce and to visualize low-dimensional views of high-dimensional data, akin to multidimensional scaling (MDS). These networks have been widely used in different signal processing and data mining applications thanks to their simplicity and performance. Nevertheless, calculating the similarity between the input vector and the weight vector using the scoring function limits their use especially when processing imperfect data. For instance, the traditional networks can not solve the case presented in the example presented. To overcome this drawback and to ameliorate the performance of these networks, we modeled the similarity by two fuzzy measures: the possibility and the necessity degrees in order to insure a wide use of Kohonen networks taking into account the imperfection of the variables (uncertain, imprecise or missing data). Briefly, these characteristics of Kohonen networks and the possibility measures can open new directions for future researches and can solve practical problems.

* **SOM Information Element Visualization**

To visualize the cluster shape and structure of a data cloud, and to achieve an efficient exploratory data analyses, several techniques have been proposed in the literature. These techniques are usually based on vector projection, using physical coordinates, color coding, etc. However, the most commonly used strategy to visualize the clusters on the SOM is distance matrices. In this technique, the distances between each unit i and the units in its neighborhood R are calculated:



The distances, or for example the median of these distances for each map unit are typically visualized using colors, although other techniques are also possible. The unified distance matrix (U-matrix) visualizes all distances between each map unit and its neighbours. This is possible due to the regular structure of the map grid. The cluster borders can be identified as “mountains” of high distances separating the “valleys” of low distances that represent the clusters themselves. It is also possible to use a unified similarity matrix. In this case, the significations are inversed. Figure 30 presents examples of both spatial and color projections as well as the U-matrix.



Figure 30

We can see that it is hard to see local details in the dense areas using the PCA, except in the interactive visualization environments where the user can zoom in on the interesting details. Contrary to PCA-projection, the map grid has equal amount of space for each map unit, and hence, map units even in the dense areas can be seen clearly.

* **Experimental Results**

In the beginning, the proposed method has been applied to the medical dataset using the batch training in which the dataset is presented entirely in each epoch. Herein, we will use the small letter p instead of O for the objects, and the capital letter P for the classes, to simplify the display upon the output units of the map. Firstly, we assume that we have a 5× 4 map and then a 15×15 output units. The learning rate and the diameter of the neighbourhood will decrease with the time of training from a maximum to a minimum value. The visualization models briefly illustrated in section will be used to display the experimental outcome. Figure 31 schematizes the unified similarity matrix using the color codes depicted in the vertical color bar at the upper left corner of figure 31. The similarity between each object and the output unites in the map are also coded with colors according to their associated bars in this figure. As it is clearly shown, the objects having the same pathology belong to the same or neighbor units. Looking closely to the coloring maps of these objects that reassemble to fingerprints, we can notice that colored maps of all the objects having the same class are approximately the same, i.e. the similar objects have similar fingerprints that differ from one pathology to another.

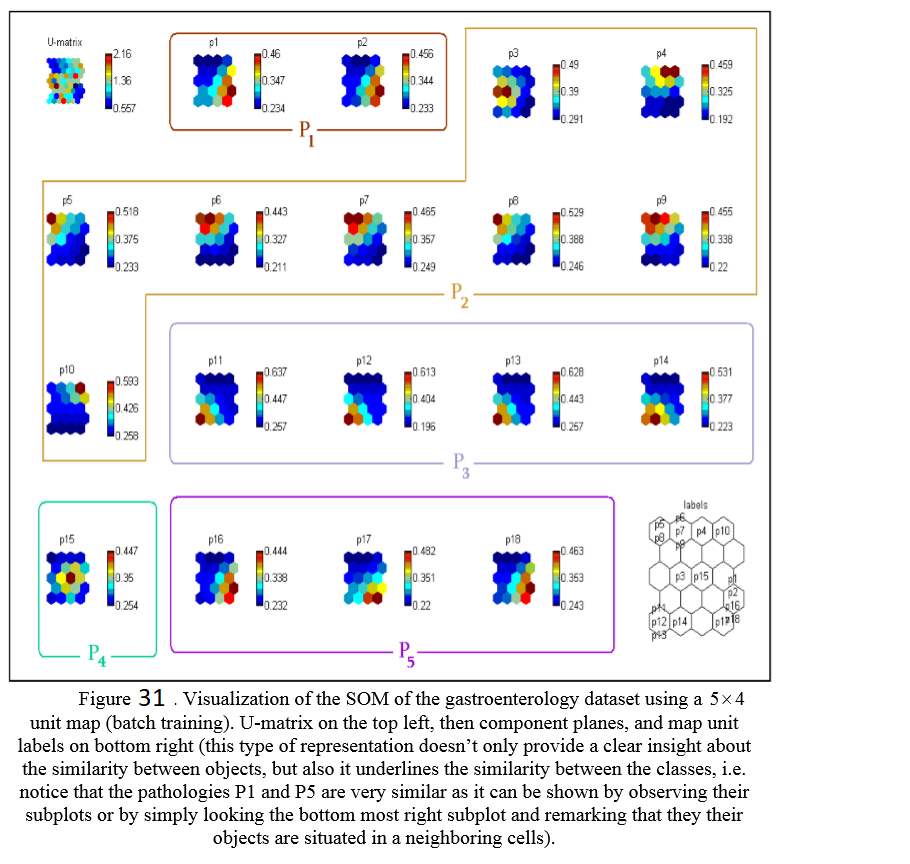


Figure 31

Figure 32 depicts another type of results representation using bar charts in each output unit that present its similarity to all the objects. Again, it can be remarked that units having objects of the same class are more similar than the other ones. The visual presentation of these information elements using the principal component analysis in a 3-dimensional coordinate space is also plotted in figure 33. Another example is given in figures 34 and 35 in which we apply the same steps to a 15×15 unit map. The same remarks and observations can be deduced. The fingerprints of each object are clearer in this example. At last, to show the validity of our approach in all the types of training, all the aforementioned steps are applied to a 5× 4 unit map using the sequential mode of training (figures 36 and 37). It is clear that we get similar analysis and conclusion from the observation of the depicted plots.

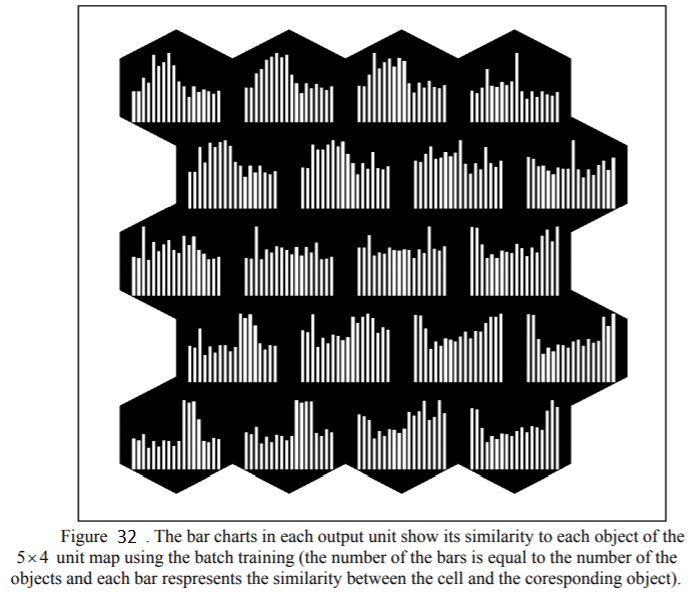


Figure 32

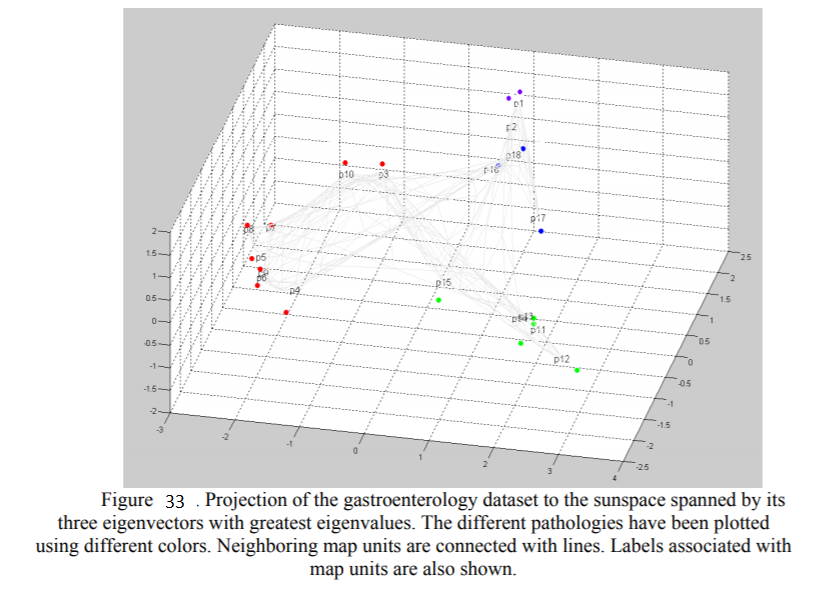


Figure 33

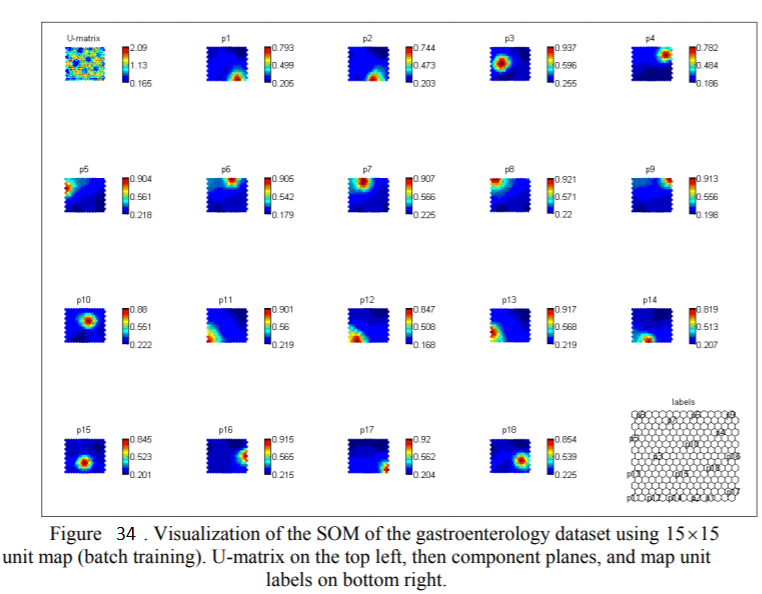


Figure 34

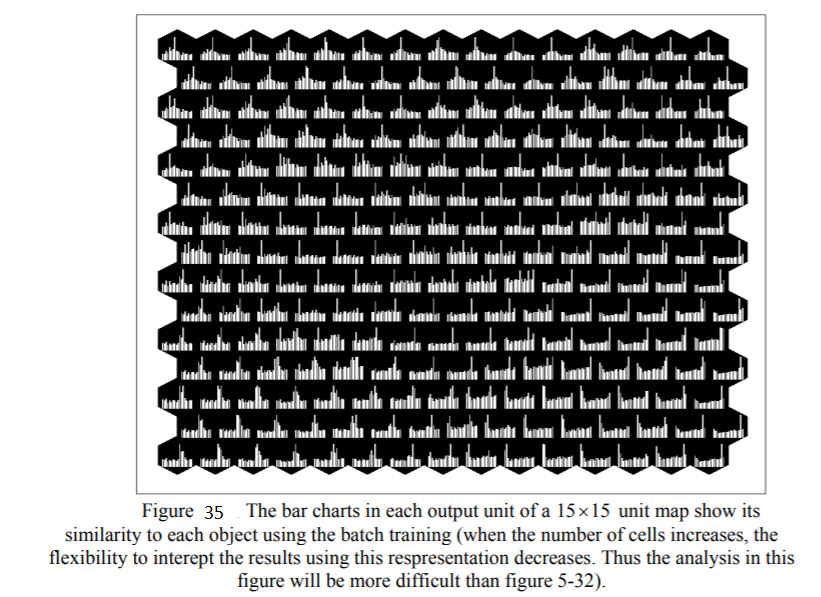


Figure 35

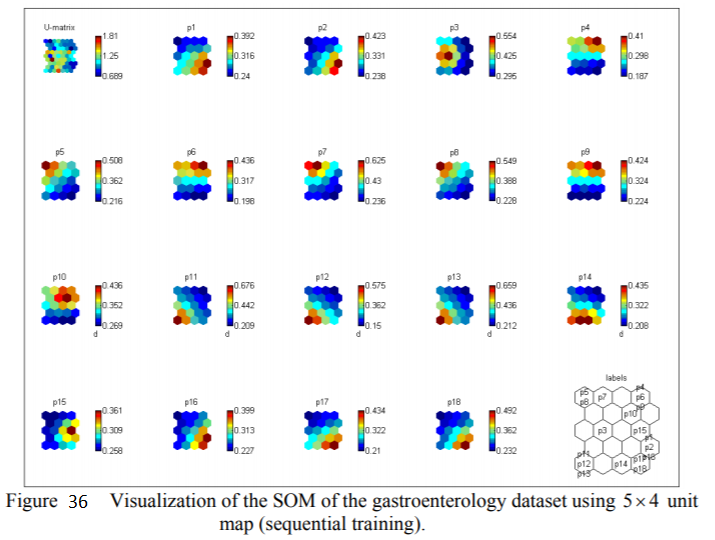


Figure 36

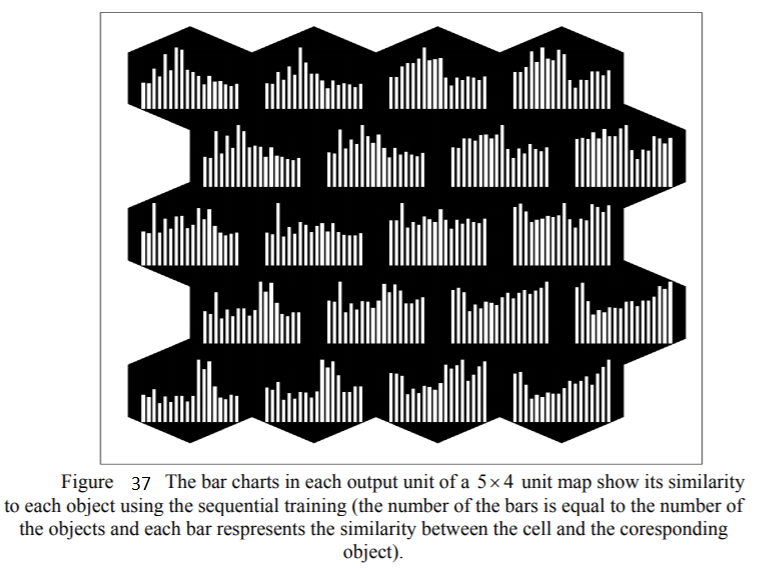


Figure 37

1. **Possibilistic Medical Image Clustering**

Advances in image acquisition and storage technology have led to a remarkable growth in image datasets size and complexity, implying an increasing difficulty for the human user to identify and to analyse underlying potentially useful unknown patterns concealed in such considerably large databases. Within the data exploratory analysis domain, clustering is one of the most fundamental tasks, having broad applications in many disciplines. It can be applied either as an independent data mining task to disclose intrinsic characteristics of data, or as a processing step preceding classification, prediction, correlation, or anomaly detection. Moreover, similarity estimation is indispensable to materialize any clustering algorithm, because a similarity operator must be applied to establish resemblance and complementary relations that may exist in data. Traditional similarity measures, are however limited to accomplish the intended objective, particularly in the presence of heterogeneous (quantitative, qualitative, ordinal) and imperfect (imprecise, uncertain, missing) data. For this reason, we propose to apply the traditional clustering techniques to the possibilistic similarity matrix proposed above, in order to take account of the heterogeneity and the imperfection of information in a unified framework. In the following, we will see some examples of applying (for the first time, to our knowledge) the evidential, the additive, the ordination-based, and the hierarchical clustering to a possibilistic necessity-based similarity matrix of imperfectly-described medical images (figure 38).

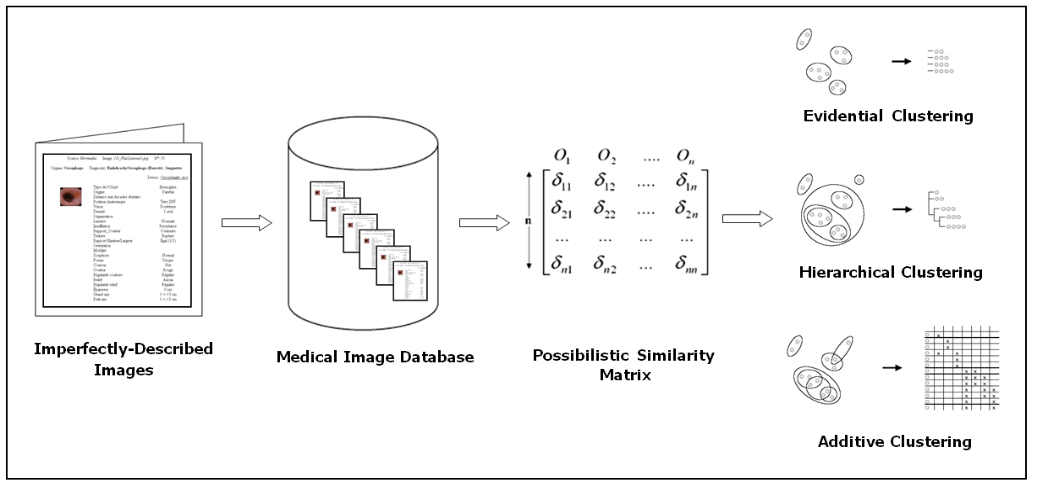


Figure 38. Clustering processing steps.

* **Possibilistic Evidential Clustering**

Given a set of  classes  and a set of  objects  where each object  from  consists of  attributes , in such a way that these attributes could be heterogeneous and could have imperfect values. We aim to assign to each element of  its appropriate class. For this purpose we propose the approach schematized in figure 39. This approach depends basically on possibility theory and on evidential clustering.

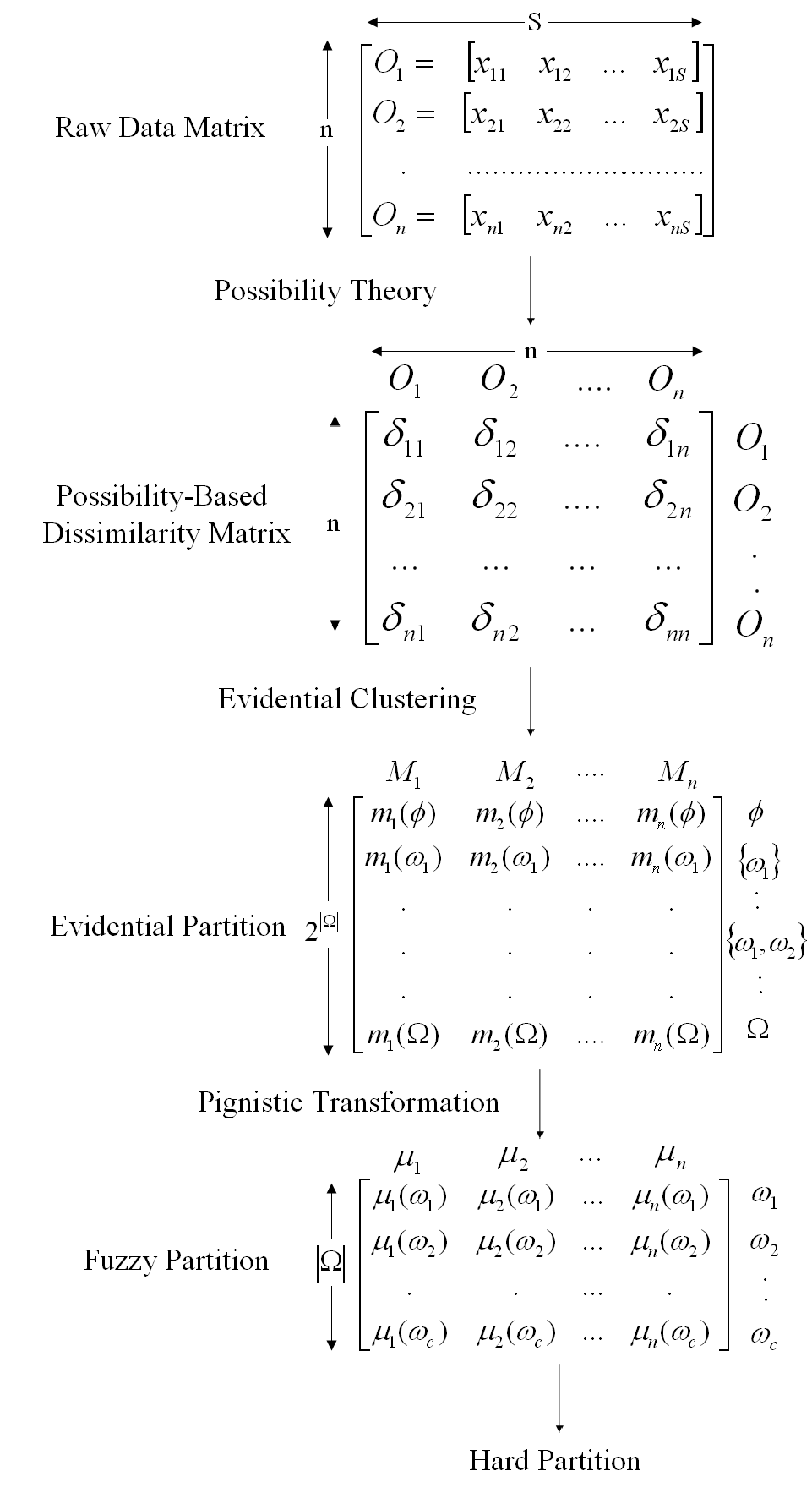


Figure 39. Possibilistic evidential clustering outlines.

* **From Raw Data Matrix to Dissimilarity Matrix**

Data to be analysed are commonly presented in one of two different formats: as a raw data matrix or as a dissimilarity matrix. The raw data matrix is an  matrix, where  denotes the value of the  variable observed for the  object. Many data mining techniques as the evidential clustering, first require the transformation of the raw data matrix into an  matrix of pairwise dissimilarities (distances) , where  denotes the dissimilarity between the  and  objects. Because of the defects and the limits of the traditional similarity measures, we propose to use the possibility theory to build the dissimilarity matrix *D*.

* **From Dissimilarity Matrix to Evidential Partition**

For two *BBAs*  and  quantifying one’s beliefs regarding the class of two objects  and , we can combine the vacuous extensions of  and  in the Cartesian product  using:

 is the *BBA* that describes one’s beliefs concerning the class membership of both objects. In , “ and  belong to the class” corresponds to the following subset of :

.

Let  be the plausibility function associated with . We have:





Given any two pairs of objects and  it’s natural that if  or equivalently: , i.e. the more dissimilar the objects, the less plausible it is that they belong to the same class and the higher is the conflict between the *BBAs*. According to Denoeux we can construct *M* easily from *D* as follows: each object  in M can be represented as a point in a dimensional space and the degree of conflict  between two *BBAs* may be seen as a form of “dissimilarity” between  and . Therefore we can transpose multidimensional scaling algorithms (MDS) to our problem, by optimizing the evidential partition *M* so that the degrees of conflict  reflect the corresponding dissimilarity. The objective function to be minimized is given by:







is the stress function used in the ordinal MDS where  is the average degree of conflict and  is any increasing function. *H(m)* is the entropy function (as we would like to extract as much information as possible from the data it is reasonable to require the *BBAs* to be as informative as possible). Actually, *H(m)* tends to be small when the mass is assigned to few focal sets with small cardinality;  is the penalization coefficient that controls the extent to which the entropy term influences the form of the solution. Increasing  for instance, will result in simpler *BBAs* with a smaller number of focal sets.

* **From Evidential Partition to Fuzzy and Hard Partition**

A fuzzy partition may be obtained from the evidential partition by calculating the Pignistic probability function  induced by each *BBA*  and interpreting it as the membership degree:

* *

A hard partition can then be easily obtained by assigning each object to the class with the highest Pignistic probability.

* **Experimental results**

The aforementioned steps are applied to the database described in appendix I. Ranks are assigned to the image descriptions, producing various intervals containing similar objects that are directly associated to the underlying examined image clusters (figure 40).

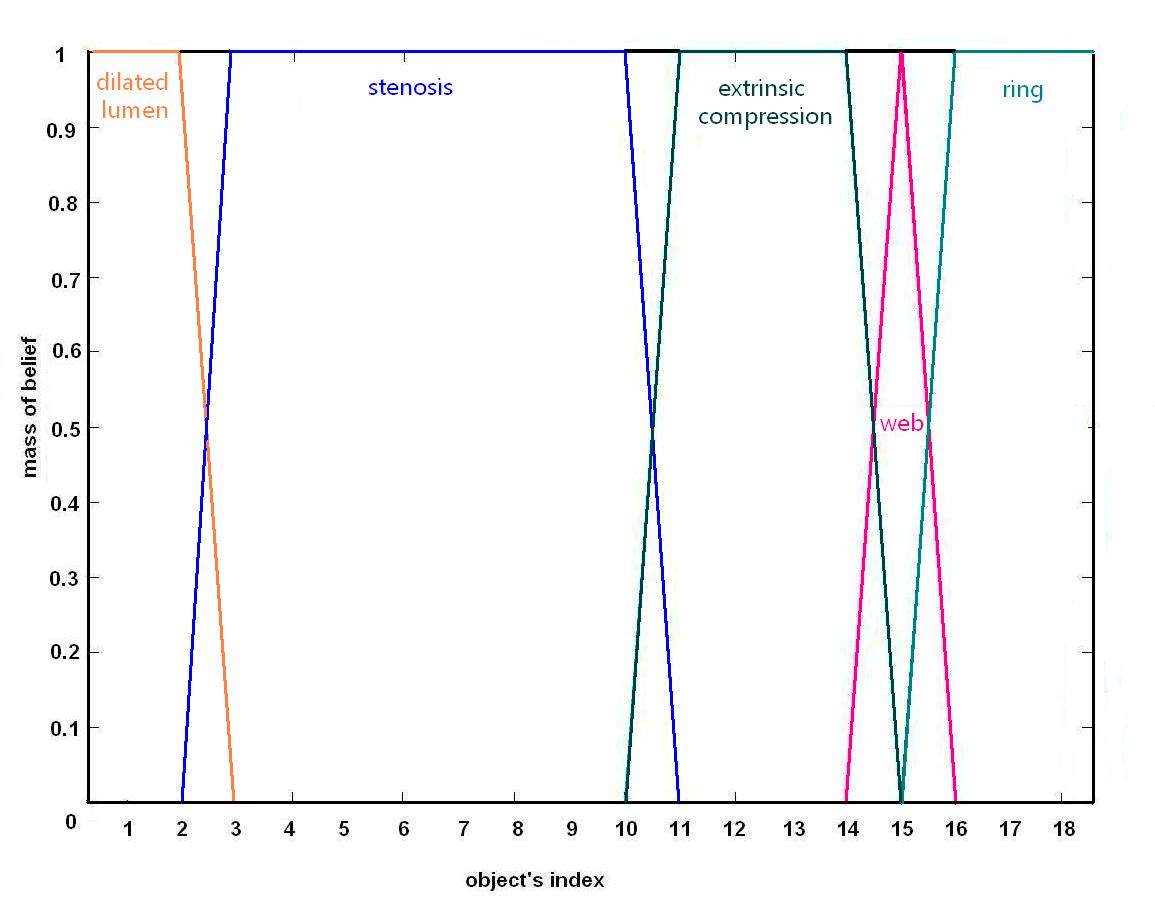


Figure 40. Evidential clustering applied to *CB*.

Formed image clusters are positioned in a sequence that reveals how many different groups of similar images were identified and which elements contain each one. For simplicity, object indexes are the same as the original sequence in the image database, but the ordering can change, preserving the same original groups. The obtained clusters match the ground truth provided by the doctor, and each object is assigned to the corresponding lesion.

* **Additive Clustering**

This conceptual approach provides a procedure for modeling the similarity relationships between objects. Given a matrix of pairwise similarities, additive clustering derives a set of weighted object clusters, which may, in various contexts, also be interpreted as domain classes. We adapted an additive clustering method focused on identification of appropriate cluster membership patterns. It uses the PBIL (Population-Based Incremental Learning) algorithm. When an additive clustering representation involving *n* objects and *m* clusters is derived, this pattern is defined by an *n  m* matrix of binary membership variables *F=[fik]nxm*, where *fik* is equal to “*1*” if object *i* is in cluster *k*, and “*0*” otherwise. The cluster weights *wk* may be found by solving the least squares problem corresponding to the Sum Squared Error (*SSE*) of the similarities:



Where the modeled similarities  are calculated as the sum of the clusters weights, to which both the *i*th and *j*th objects belong:



The similarity of two objects is modeled therefore as a weighted additive measure of their common clusters. Like conventional clustering models, additive clustering recovers discrete subsets of objects weighted by salience, for which the similarity increases proportionally to the number and weight of their common clusters. However, instead of applying it to a similarity matrix, our approach is based on a possibilistic matrix.

Similarity is represented by the weights assigned to the clusters and the labelled members of each cluster (figure 41).

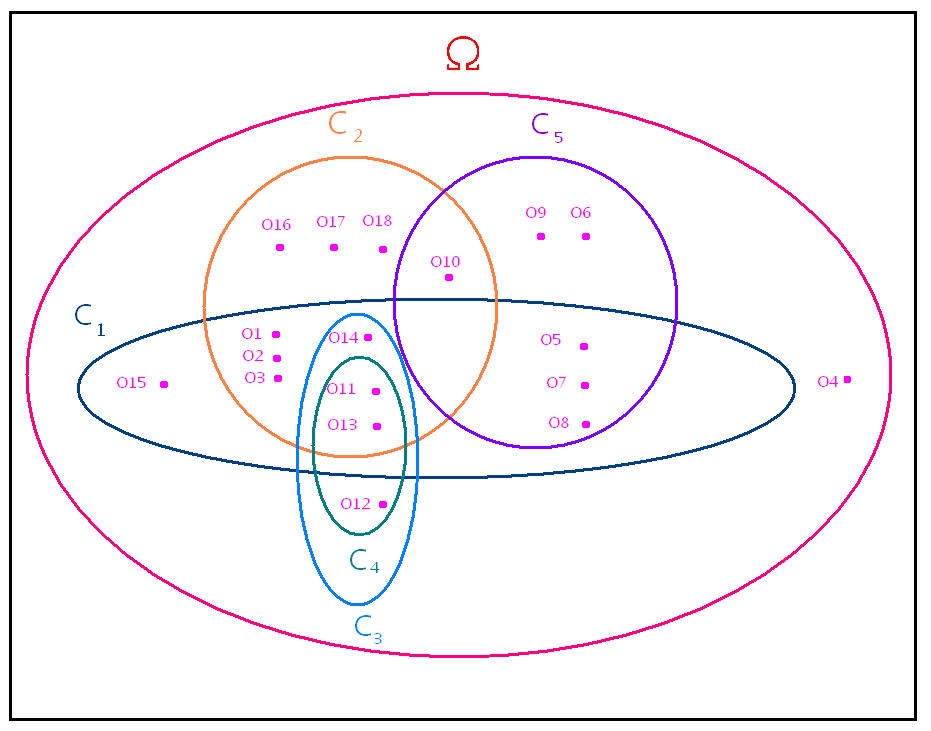


Figure 41. Additive clustering applied to *CB*: , , , , , .

Clustering is visualised through a pattern of points that represents the database images (*On*) *n=1, 2, ..., 18*, distributed according to the degree of similarity between them. Five classes, *Cl*, *l*=1, 2, ..., 5, and a class for the complete ignorance *CΩ*, or rejected class, have been identified according to the strength of belonging to a class, provided by the calculated weights. Similarity between any pair of images increases as the common class numbers and weights increase. The fact that one image is included in several classes, indicates that it belongs to them with different degrees of importance, given by the corresponding weights. Comparatively, overlapping clusters also designate groups of images which are more similar than others. Furthermore, image clusters per pathology can be identified calculating the respective similaritie. Even though results correspond to ground truth, their interpretation is somewhat tricky, requiring specialized algorithm knowledge.

* **Ordination-Based Clustering**

Ordination, sometimes called seriation or sequencing, is a frequent issue in data mining that aims to arrange all objects in a set in a linear order with the objective of revealing underlying structural information in such a way that we can visualize simple geometrical and relational structures between objects to explain the pattern of information potentially present in a numerically given dissimilarity matrix. Specifically, to order a set of  objects , one typically starts with an  symmetric dissimilarity matrix  (where  for ,  represents the dissimilarity between objects  and , and  for all ) and a permutation function  (a function which reorders the objects in  by simultaneously permuting the rows and the columns of the dissimilarity matrix). The goal in the ordination is to find a permutation function  which optimizes the value of a given loss function  (or a merit function M) as follows:





leading to obtain a anti-Robinsonian dissimilarity matrix  in which the small values of dissimilarity (large values of similarity) are concentrated around the main diagonal as closely as possible, whereas large values of dissimilarity (small values of similarity) fall as far as possible from it. In other words, the matrix  respects the two following gradient properties:

 for  (for the rows)

 for  (for the columns)

* **The Loss and Merit Functions**

Many loss and merit functions have been proposed in the ordination literature such as:

1- The loss function proposed by Chen to quantify the divergence of a matrix from the anti-Robinsonian form given by:



Where 

 is a function which defines how a violation of the two aforementioned gradient properties for an object triple (, and ) is counted.  is an indicator function returning “1” for the violations.

2- The loss function proposed by Caraux to quantify the deviations between the dissimilarity in  and the rank differences of the objects defined by:



Where  is the rank difference or gap between  and .

3- The Hamiltonian path length loss function proposed to optimize the ordination with respect to dissimilarities between neighboring objects that constitute the vertices in a weighted graph:



4- The merit function proposed by Hubert to quantify the divergence of a matrix from the anti-Robinsonian form:



Where  is a function which defines how a violation or satisfaction of the gradient properties for an object triple ( and ) is counted:



5- The merit function proposed by Caraux to represent the moment of inertia of dissimilarity values around the diagonal as:



6- The merit function proposed by McCormick called “the measure of effectiveness”:



In fact, any function from this list (the six proposed functions) can be used without any preference, because they are all simple and easy to be optimized. It has to be noted also that  can be brought into a perfect anti-Robinsonian form by row and column permutation whenever  is an ultrametric ( for all  (the three-point condition), and  when ) [2]. However, for most data only an approximation to the anti-Robinsonian form is possible. Hubert [2] shows that a best fitting ultrametric, say  to a given dissimilarity matrix  can be generated by applying the iterative projection strategy of Dykstra to find an optimal solution to the system defined by the ultrametric matrix constraints and by minimizing  given as follows:







Thus, the anti-Robinsonian matrix is obtained by applying the permutation function on :





Where 

In our experimental part,  is calculated using the aforementioned manner, for two reasons: on the one hand, since  is anti-Robinsonian, it can be used in revealing the potential structure of the set of objects (ordination) [1], on the other hand, since  is a reordered ultrametric, it can be very useful in obtaining hierarchical partitions and tree representation of the set of objects.

As the other techniques of data mining, the construction of meaningful and robust dissimilarity matrix is essential to reveal the structure of the data. Therefore we propose to use the possibilistic similarity that enables us to estimate the proximity between objects having heterogeneous types of attributes (quantitative, qualitative, ordinal, distribution …etc) and imperfect data (missing, imprecise, and/or uncertain data).

* **Ordination-Based Clustering**

A similarity matrix containing the necessity degrees of similarity between all the objects of the tested database (described in the appendix) is constructed and transformed to a dissimilarity matrix by applying decreasing function. We have chosen a graph-based model that depends on the anti-Robinsonian form of the matrix to depict a tree representing the hierarchical partitions of the database. This has been achieved by fitting an ultrametric matrix to the constructed dissimilarity matrix and then by reordering the ultrametric matrix using the permutation function given Hurbert’s equation (the forth case in the precedent section). We have remarked that the ordination has been carried out in such a way that all objects belonging to a given type of pathology lie next to each other, and a group of similar objects belonging to a given pathology lies always next to the group of the pathology the most similar. We have to underline that we can also distinguish thanks to this approach the similarity degrees between the objects belonging to the same pathology class. Figure 34 depicts the anti-Robinsonian matrix of the dataset  (the reordered ultrametric-fitted dissimilarity matrix using the row and column order of ). Here the blocks of equal-valued entries are highlighted with the same color, indicating the partition hierarchy induced by the ultrametric (depicted in table 14 which is a clearer equivalent version of figure 42). Note that at the beginning each object represents an individual class, then at the first level 0.3987 (the smallest value) the two most similar classes merge (agglomerate) together to form one class. The agglomeration continues gradually until all the objects gather together in just one class (at the highest level 0.7540) which represents our entire dataset. For the partition hierarchy just given, the alternative structure of a tree (dendrogram) for its representation is given in figure 43 (the terminal nodes (vertices) of this structure correspond to the 14 objects of ). Figure 44 gives another example of a dendrogram induced by the ordination applied to the database . As expected, we can see that objects having the same class lie next to each other in the resulting ordination. Other interesting remarks and interpretations can be obtained: for example, we can say that  and  are the objects the most similar in this base (actually, they are the objects the most similar in comparison with all the other objects). We can also see that the pathology  is an extinct pathology because it can be distinguished easily from the other classes of pathologies. We can also remark that object  is the object the less similar to the other members of the class .

According to the levels of the members of a given class, we can divide this class into smaller homogeneous groups (subclasses) in order to depict the details and the degrees of a given pathology (for example ulcer grade I, grade II, and grade III). Object segmentation and organization in this way can be very useful for doctors to find similar cases (content-based case retrieval) or to help to take a decision or to find a solution (diagnostic) to a similar case (case-based reasoning). It helps also to study the association and the relations that exist between the pathologies and the lesions on the one hand, and those that exist between the objects composing the same pathology on the other. In fact, this representative model provides the doctors with a simple-to-interpret tree that visualizes explicitly the relations between the objects and the pathologies of our database.

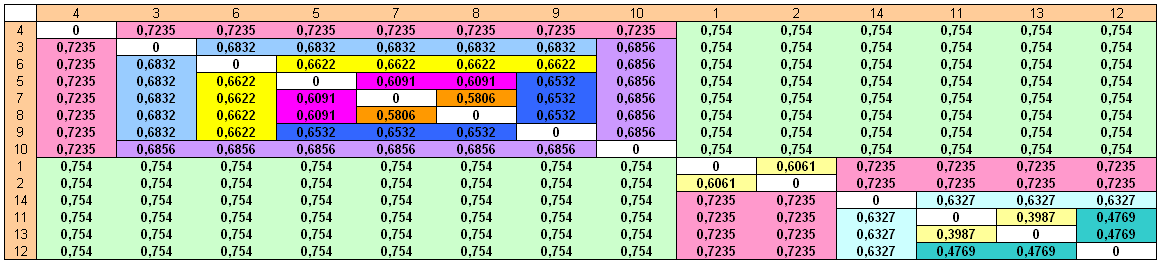
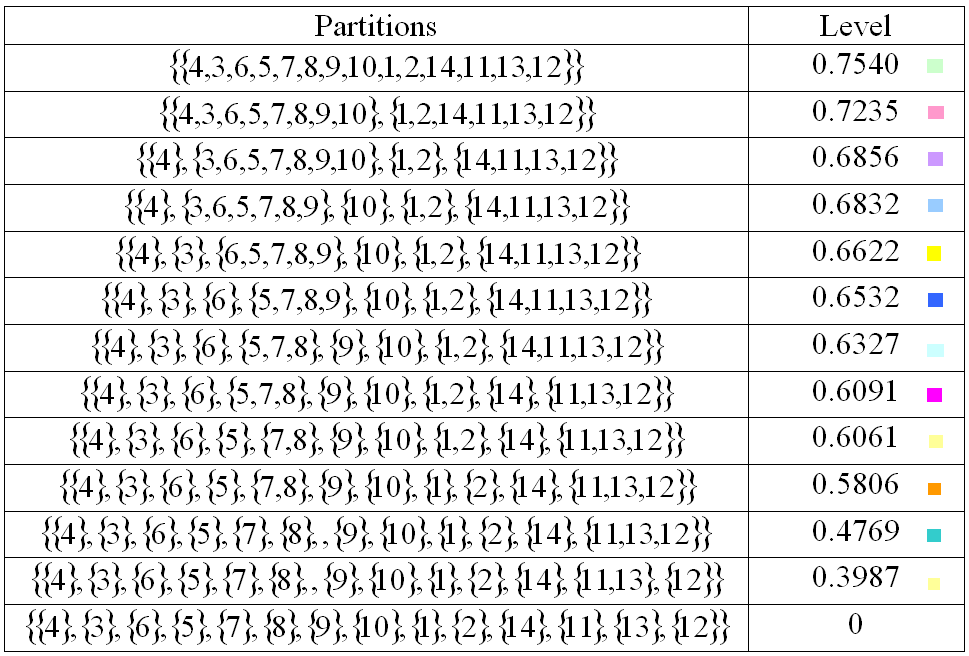


Figure 42 .The anti-Robinsonian matrix of *CB1*.

Table 14. The partition hierarchy induced by the ultrametric anti-Robinsonian matrix (note that the objects that represent a given pathology tend to be agglomerate in the same cluster).



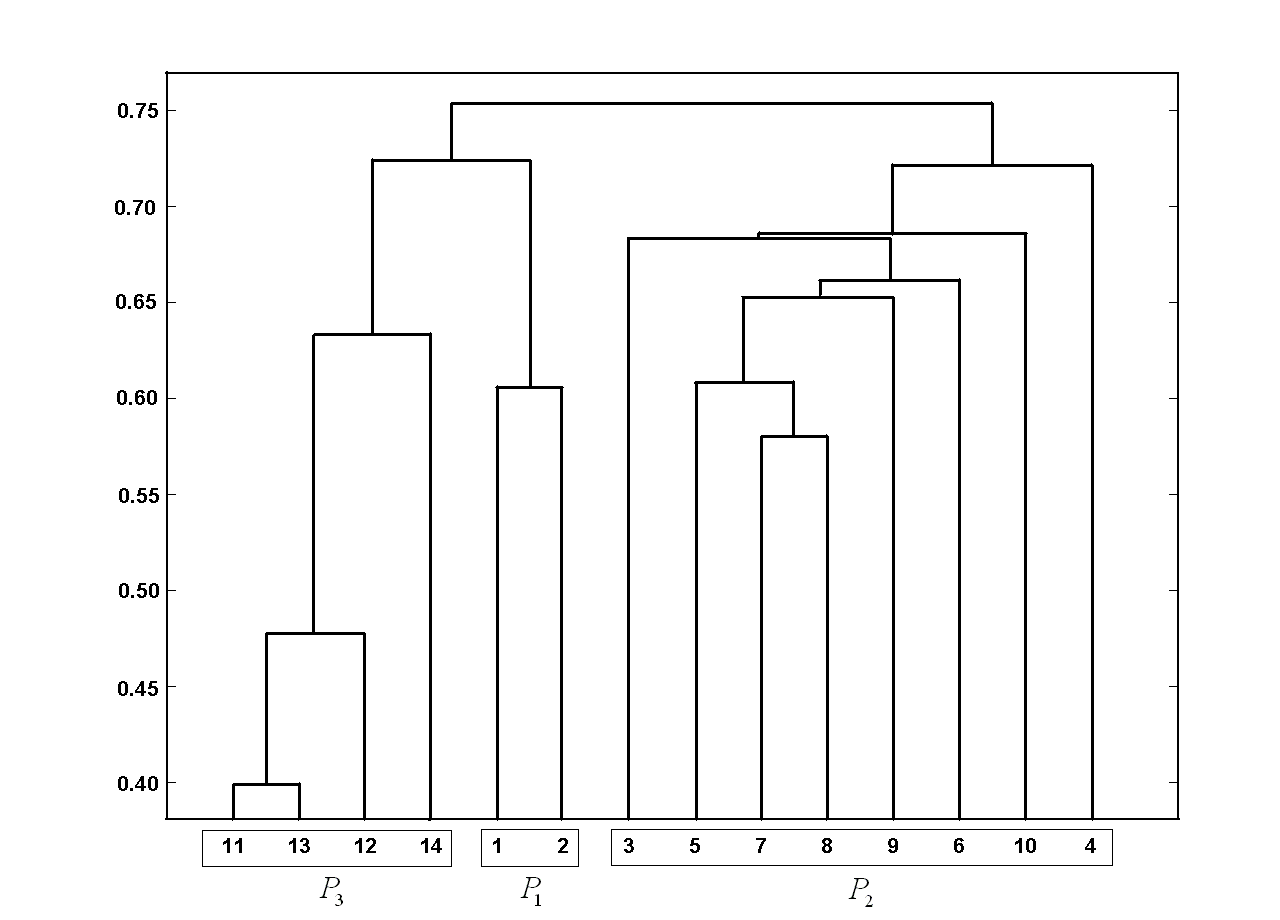


Figure 43. The dendrogram of *CB1*.

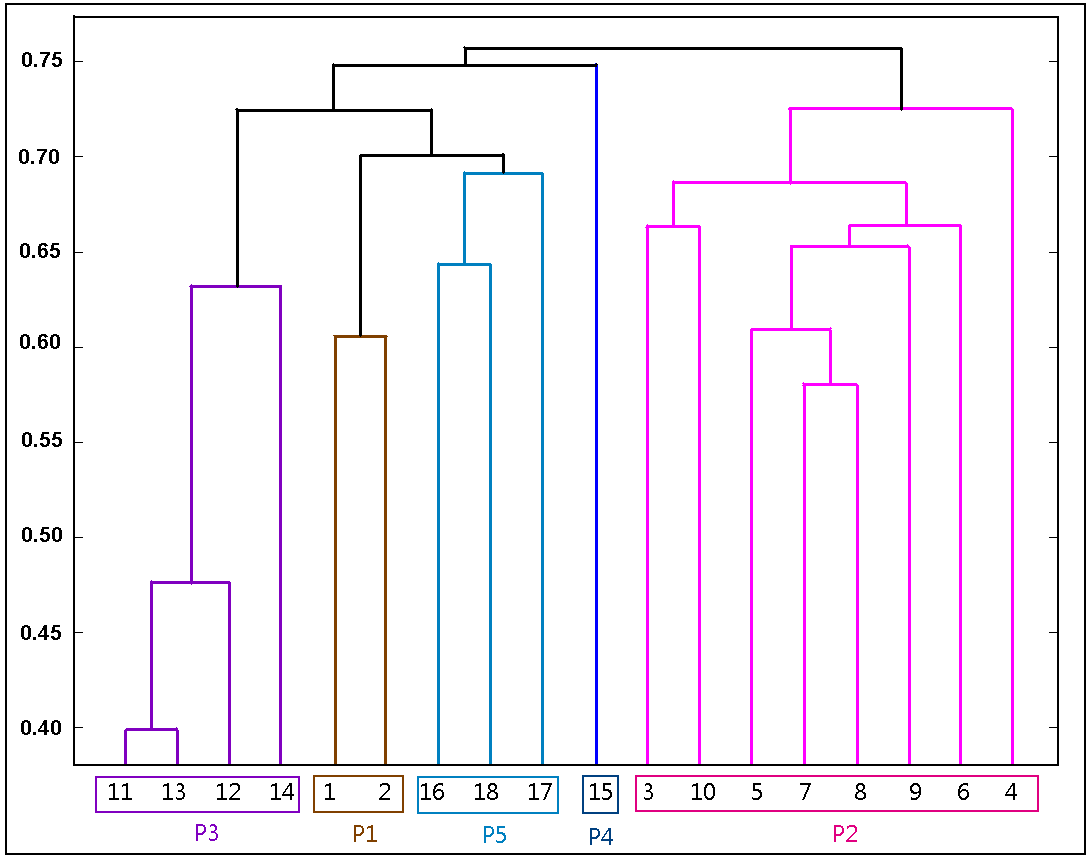


Figure 44. The dendrogram of *S* generated by the anti-Robinsonian matrix.

In brief, ordination is the task of optimally sequencing (or seriating) a set of objects along a continuum, based on the available dissimilarity matrix between the object pairs with the goal to reveal and to explain the relationships that may be present among the objects, as reflected by the proximity data. This process can be realized by constructing a best reordering for the rows and simultaneously the columns of ultrametric dissimilarity matrix. The row/column reordering to be identified optimizes, over all possible row/column reorderings, some specified measures of patterning for the entries of the reordered matrix. A variety of specific indices of patterning has been presented above. These indices aim to construct an anti-Robinsonian matrix of the data under construction. This matrix can be presented by a dendrogram and can be easily interpreted to reveal the structural pattern in the database. To our knowledge, this is the first time that the ordination techniques are applied to a possibilistic matrix. The obtained results show the clustering of the images has be carried out as expected, and the obtained clusters correspond to the different pathologies presented in the test base.

**10-5- Possibilistic Hierarchical Clustering**

The dendrogram is one of the powerful tools particularly used in hierarchical clustering to visually explore and summarize the correlations between the patterns of a given dataset. As we have seen in before, hierarchical clustering algorithms create nested sets of ordered objects, depicted by the dendrogram, on which each node represents a cluster. While, the root node includes all the individual patterns, each descendant contains a subcluster of its parent node. The height of each node is proportional to the measure of similarity between the two concerned sub clusters, and the leaf nodes typically represent individual patterns.

* **Hierarchical Clustering Experimental Results**

We calculated the possibilistic similarity of all the pairs of objects represented as imprecisely-described digestive images in the dataset described in appendix. Then, we applied the techniques presented before to this possibilistic matrix using different types of linkages (inter-class similarity measures with the aim to display the possibilistic estimated similarity matrix to reveal the potential patterns and to detect the outliers of this set. As expected (shown in figures 45), in all the cases, we obtain representational models compatible with the classes of pathologies assigned by the expert (the truth ground). When we have only one example of some pathologies like the object *O15*, this object (example) will be detected as an outlier, in addition to the object *O15*, *O4* detected as an outlier by all the other perspectives of visualizations. We point out that pair (O11 and O13) contains the objects the most similar in the entire tested dataset as anticipated (this fact is also indicated in the truth ground).

|  |  |
| --- | --- |
| SINGLE2  (a) Single linkage | complete  (b) Complete linkage |
| average  (c) Average linkage | ward  (d) Ward linkage |

Figure 45. The dendrogram of S using different types of linkage.

In brief, our work here consists of three main parts. The first part concerns the measuring of the similarity among the images of our dataset in spite of the imperfection of the heterogeneous describing features by means of the possibility theory measures (the possibility and the necessity degrees of similarity). The second part consists of exploiting this similarity modeled by the necessity degrees in several clustering techniques to form the essential groundwork for the last phase whose goal is to discover the potential structure and patterns in the image set (image exploratory analysis and mining). As we have seen, clustering analysis is a very powerful method in visualizing the abstract conception of similarity to understand and to interpret the observed events in a meaningful manner. In the additive clustering applied to our data set, the similarity is represented by the weights assigned to the clusters and the members of each cluster as given in equation. In the hierarchical clustering and the ordination-based algorithms, the similarity between any two objects is presented by the length of the path that connects these objects (images) represented by the terminal nodes in the ultrametric tree constructed. In the evidential clustering, similarity can be represented by the basic belief assignments and by all the partitions resulting from it (possibilistic, fuzzy, and crisp partitions). The results obtained in these experiments show that a strong similarity exists among the images that have the same type of lesions. In other words, an image belonging to a given lesion’s class is more similar to any other image from the same class than to the other images belonging to the other types of lesions. Without any a priori knowledge of the key attributes of the studied lesions, and in spite of the imperfection and the heterogeneity of the attributes’ values, we were capable to organize and to classify the image database in a meaningful manner. As we see, the hierarchical representation gives clearer similarity visualization (representation) for the end user than the evidential and the additive clustering, inasmuch as we need to make some more mental analysis and calculation in order to understand the resemblance among objects in the two last clustering. Besides to its obvious comprehensible representation and unlike the other two clustering techniques, the hierarchical analysis does not need to know in advance the number of the clusters (pathologies) which is sometimes impossible because we can not have always a priori knowledge of the treated dataset. Instead, it gives us all the possible collections of clusters in all the similarity levels. Nevertheless, we can not confirm that it is the best representative model of similarity in all the cases, because we need sometimes in some applications to combine several similarity points of view concerning a set of images described by several experts. In this case, applying the evidential clustering is better, because it represents the similarity as belief masses and combining these basic belief assignments is very easy thanks to the already established methods in data fusion (like the methods of Dampster-Shafer, Yager, and Dubois-Prade). Furthermore, the evidential clustering gives deeper insights of the processed images especially in the cases where two different types of lesions exist in the same image. On the other hand, using the additive clustering in machine learning and artificial intelligence in some applications could be very appropriate since its structure is very alike to the artificial neural networks.

1. **Appendix : Image Database Description**

The possibility-based similarity measuring and modeling is tested along this paper on a digestive endoscopic atlas of documented endoscopic lesion descriptions, and scene information of the upper gastrointestinal tract, esophagus, stomach, and duodenum. Database image attribute descriptions characterize observed anomalies or lesions identified by an expert, according to a well-defined and exhaustive description structure:

* Object location: anatomic (longitudinal), position in the organ (axial), and distance from the teeth.
* Repeated objects: number of identical objects and spatial organization.
* Object aspect: shape and edge, dominant color and color regularity, relief and regularity of relief, sizes (major, minor axes and thickness), axes ratio and major axis orientation, height, motility, effect of insufflations, and consistency.
* Relation with adjacent organs: color contrast, texture contrast, and consequences on the lumen.

These descriptors represent 24 features, summing 145 distinct values for simple objects. Attributes are either semantically or numerically coded, with the help of an adapted interface by an expert physician (figure I-1). Complex objects are defined when two or more simple objects are related on the same visual scene. Each one has its own attributes and a spatial relation (closeness and order, rated as: into, in contact, in contact and upward, around, and around and upward) links them, resulting in 33 features and 206 different values (figures I-2 and I-3). Depending on the secondary objects types, other relationships may appear like relative sizes and consistency, combined with the absence of some features in some objects, uncertain and incomplete descriptions.

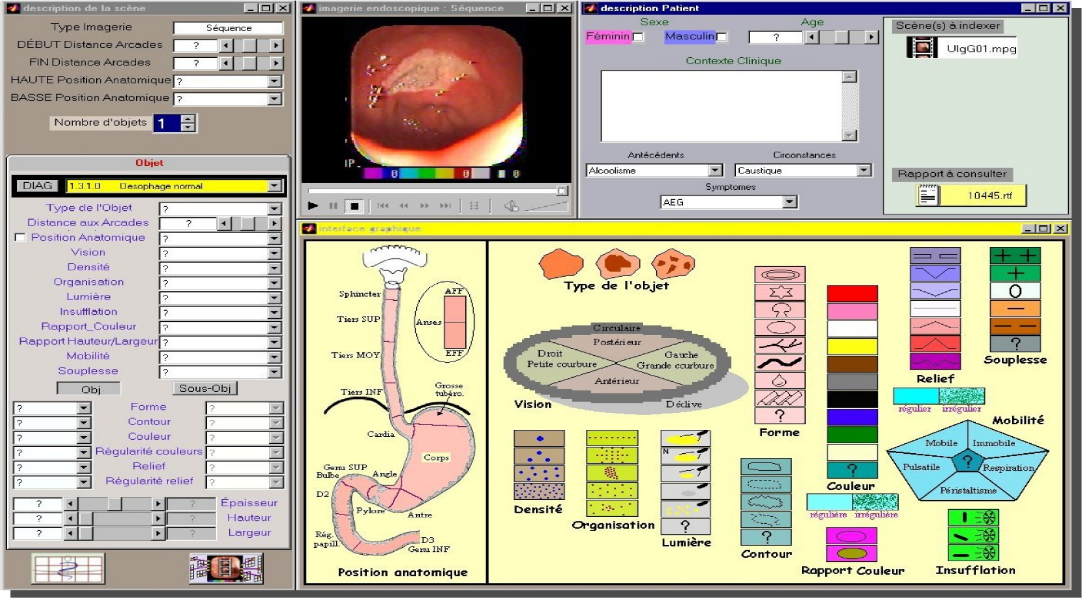


Figure I-1. The interface used to assign the values of the attributes of the dataset.

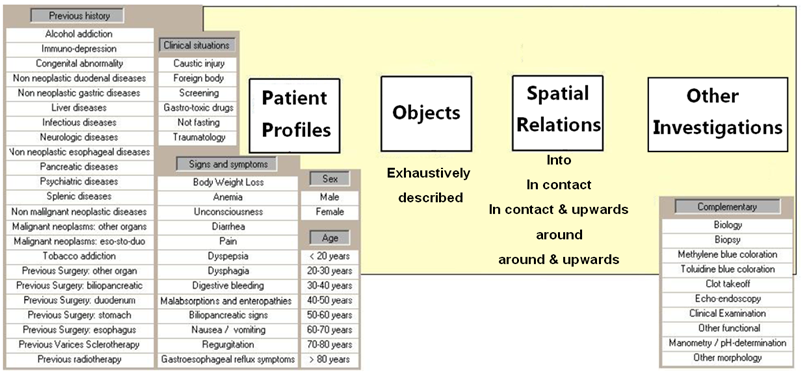


Figure I-2. An example of the object structure.

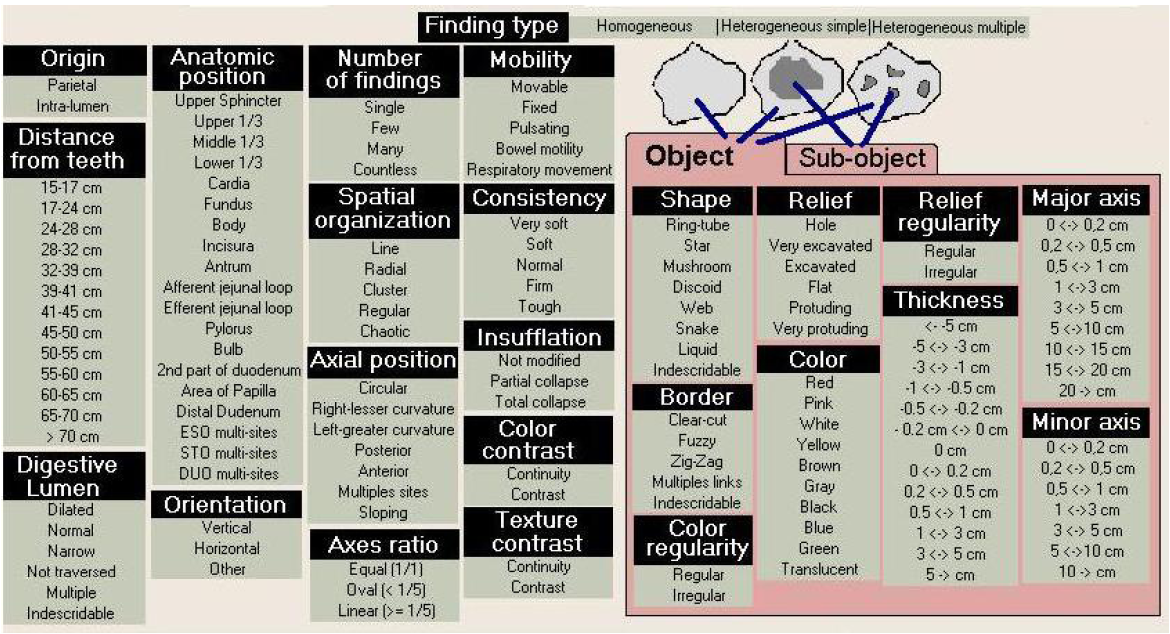


Figure I-3. An example of the attributes used in this dataset.

Endoscopic diagnosis relies on the analysis of associations between these elementary lesions and the medical context, which includes sex, age, clinical antecedents, consultation circumstances, and symptoms. Complementary exams may be necessary to determine whether the initial diagnosis is confirmed or refused. These exams include histological examination of biopsy specimens, coloration of the digestive mucosa, and morphological or functional evaluations. Once processed, all these elements and diagnostic decisions are then recorded on a medical report. Among the documented lesions and pathologies we find: dilated lumen, stenosis, extrinsic compression, web, ring, hiatal hernia, undigested food, liquid blood, blood clot, z-line, spot, circular Barrett’s, Moniliasis, simple erosion, ulcer, and Petechial mucosa.

Lesions and diagnosis are intended to be independently described under this scheme, even though practical experience shows that endoscopic findings may point towards a particular diagnosis, whereas other diagnosis alternatives including the same lesions could also be specified. For this reason, the digestive endoscopic atlas also defines specialized findings, which are classes that describe more in detail the generic diagnosis. Validated by medical experts, the atlas consists of 89 endoscopic findings, 126 endoscopic diagnoses, and 118 specialized findings, a priori descriptions.

To illustrate our results in a clear and simple manner (especially when the results are visually displayed), a subset of cases belonging to the main image database were processed, given that the respective values correspond to a sub matrix of the main similarity matrix, and do not change if originally calculated from it. This facilitates understanding the studied possibilistic approaches, applications, results, and the graphical representation, and emphasizes the general character of the proposed approach. Defined as *CB = {O1, O2, …, O18}*, the subset of cases contains 18 described images, presented in figure I-4. *CB* is structured in the following manner according to the ground-truth provided by the specialist: *P1 = {O1, O2}* corresponds to the “Dilated Lumen” pathology; *P2 = {O3, O4, O5, O6, O7, O8, O9, O10}* conforms with the description of “Esophagus Stenosis”; *P3 = {O11, O12, O13, O14}* is a set of images that represent the “Extrinsic Compression” pathology; *P4 = {O15}* describes the “Web-Shape” pathology; *P5 = {O16, O17, O18}* is a set of images on which the “Ring-Shape” pathology is visible. Some examples of this contribution will be applied firstly to the dataset ****, defined as:*CB = {O1, O2, …, O14}* in order to assure the clarification of come delicate application.

|  |  |  |
| --- | --- | --- |
| *P1* | Dilatation02 | Dilatation01 |
| *P2* | 3C_Muc_Infl_refluxEsophagitisG4_3 | 15E_Muc_Infl_refluxEsophagitisG4_5 |
| 17C_Muc_Infl_refluxEsophagitisG4_4 | 15B_Muc_Infl_refluxEsophagitisG4_1 |
| 11D_Muc_Infl_BarrEsop_C7 | 7B_Muc_Infl_BarrEsop_C6 |
| 7D_Muc_Infl_refluxEsophagitisG2_2 | 2E_Protusions_tumor_infiltrative4 |
| *P3* | 4E_Protusions_TumorSubmucosal5 | 18B_Protusions_TumorSubmucosal1 |
| 4E_Protusions_TumorSubmucosal5 | 15A_Protusions_TumorSubmucosal3 |
| *P4* | Diaphragme01 |  |
| *P5* | Schatzki03 | Schatzki02 |
| Schatzki01 |  |

Figure I-4. Cases sample of the tested image database.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_