

Data Representativeness Based on Fuzzy Set Theory

Frédéric Blanchard, Philippe Vautrot, Herman Akdag, Michel Herbin

▶ **To cite this version:**

Frédéric Blanchard, Philippe Vautrot, Herman Akdag, Michel Herbin. Data Representativeness Based on Fuzzy Set Theory. Journal of Uncertain Systems, 2010, 4 (3), pp.216-228. hal-00694895

HAL Id: hal-00694895
https://hal.science/hal-00694895

Submitted on 7 May 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Data Representativeness Based on Fuzzy Set Theory

Frédéric Blanchard*†, Philippe Vautrot, Herman Akdag and Michel Herbin

Abstract

This paper presents an original definition of *data representativeness*. The representativeness of each datum in a dataset is a meaningful notion quantified by a degree computed by aggregating fuzzy subsets. These fuzzy subsets are obtained by fuzzifying data in a robust way. We illustrate the usefulness of the representativeness by presenting applications for statistical location estimation,and for cluster analysis.

Keywords : Data Representativeness, Fuzzy Set Theory, Robustness, Data Analysis, Clustering, Multidimensional data, Multivariate Analysis, Ordinal Analysis, Rank Transformation.

1 Introduction

In the past few years the growing computing capabilities and the improvement of data acquisition have led to the need for managing new kinds of data. The extraction of information from datasets is a crucial task. Data-mining and data analysis theories offer methods and techniques that allow exploration, description and explanation of these datasets.

In the context of statistical theory data sets are generally considered as representative cross-sections of theoretical populations. More precisely, most of the involved tools consist in estimating the parameters of a hypothetical underlying distribution of the sample. Classical instances of location estimators are the *median* and the *mean*. The *standard deviation* is a common dispersion estimator. These tools predominantly study the probability distributions and are most often based on assumptions about them. We assume that classical statistics are often not representative of datasets.

The lack of data is another critical point. Data usually contain intrinsic uncertainty or imprecision induced by the acquisition, the nature of the data etc. In addition, real datasets often contain outliers and noisy data. Finally, it

*CReSTIC, Rue des Crayères, BP 1035, 51687 REIMS Cedex 2, FRANCE
†Contact : frederic.blanchard@univ-reims.fr

is usually impossible to make the assumptions on the data distributions needed by classical statistical and probability tools [17],[5].
To deal with these problems, some efficient concepts can be used. The theory of fuzzy sets and the possibilistic approach, are common ways to deal with imprecision and uncertainty. As said before, the sensivity to outliers is an element that often leads the most common statistical methods to fail. The concept of robustness, inherent to the non-parametric methods, offers "resistance" to outliers. More generally, non-parametric concepts like ranks and order statistics [19],[35],[38] permit both to beat the curse of outliers and noisy data, and need no assumption on the data distributions.

Our goal is to propose a method that extracts information from data relatively to the given dataset. The proposed tool has to be as insensitive as possible to outliers and to deal with lack of data. As an important constraint, we wanted to develop a non-parametric method needing no assumption on the distribution of data [10],[17].
To address this question we have defined a new theoretical notion expressing the representativeness of data, relative to a given dataset. This new notion affects to each datum a quantitative information that expresses how this datum is representative in the dataset. In other words, it can be viewed as a way to answer the question: "*How* reprentative is one datum in his sample ?". By adapting the semantics, we introduce different applications of this notion. By searching the "best representant" in a sample we define a new location estimator of the underlying distribution. By finding some best representant of subgroup, and linking data with these "centers", we propose an algorithm for cluster analysis.
How the representativeness is constructed ? Our concept of representativeness is defined by using different efficient and eprouved concepts. The goal of our methodology is to compute a degree of representativeness for each datum of a dataset. This degree is computed as the aggregation of fuzzy subsets [41] associated to data. The first step of our method consists in fuzzifying data [42],[7]. This fuzzification is not classical. We choose a robust technique that uses first a rank transformation (see [36],[30],[23]) of the dissimilarity between data. This technique makes our method free of assumption about the distribution [17],[10]. This way is really different from a classical fuzzy ranking or fuzzy ordering approach [16],[1]. The aggregation is then realized using an OWA operator defined by Yager in [41]. This aggregation operator brings flexibility and permits to attenuate the effect of outliers. Finally, we obtain for each element of the initial dataset a value that quantifies how this element is representative in his set. In this approach, the computation of representativeness is made exclusively on the finite input dataset and not in an hypothetic underlying space. Although we are not concerned by the problems of sampling, estimation or interpolation.

To illustrate the interest of our contribution we present its use in different theoretical contexts. First we can use the representativeness as a location estimator. The simulations and the experimentations proove that our induced statistic is efficient and gives better results than the median and other classical

statistics. Its efficiency with different kind of distributions and in particular with assymetric distributions is presented. Secondly we illustrate the use of the representativeness for data clustering [22],[27],[32]. Our concept permits to link data and to define clusters. One advantage of our approach is that the separability between clusters is not assumed to be linear.

The next section of this paper presents the theoretical basis and definitions of our contribution. In section 3, we illustrate the use of the notion in two major applications: statistical description of a sample, and data clustering. Finally we discuss the different choices we made and propose a conclusion for this paper.

2 Theory

We decribe in this section the theoretic basis of our concept which is based on the fuzzy set theory.

The concept of fuzziness was introduced in 1965 [42] to deal with imprecision and vagueness. It allows for instance to represent and process imprecise statements like "Bob is young", "Alice is tall", or rules like "if an obstacle is close, then brake is immediately".

In a data analysis problem, the nature of data depends on the domain of application. We can have to analyze financial data, spatial data, or image data for instance. In most of cases there is an intrinsic imprecision induced by the acquisition technique. For example, in the case of image data, there is an imprecision caused by the limitations of the captors. According to the resolution of a captor, each pixel of a photography matches to different area's size of the captured scene. Therefore, the information contained in each pixel could be disturbed and photos are sometimes blured or noisy. For each kind of data, it's necessary to consider and to manage imprecision and uncertainty. Fuzzy sets and possibiliy theory are a common and efficient way to deal with this problem.

We introduce in this paper a notion of *fuzziness of data relatively to its dataset*. This notion allows us to define a kind of fuzziness for data, depending on the attributes of the dataset. This fuzzification step leads then to the definition of *data representativeness in a dataset*.

2.1 Data as fuzzy sets : a robust representation

As said before, real data are usually imprecise and the observations in datasets can be subsequently represented as fuzzy sets.

Let Ω be a dataset formed by n observations. $\Omega = \{x_1, x_2, \ldots, x_n\}$. We decide to represent each x_i (for each $i \in I = \{1..n\}$) as a Fuzzy subset. We note

$\tilde{x}_i$ the corresponding fuzzy set.
The membership function associated to each fuzzy set $\tilde{x}_i$ is not defined as usual[7],[4]. We chose to define these membership function with the intention of dealing with outliers and with datasets where data are distributed according to various and heterogenous underlying probability distributions.
In clustering problems [11],[20],[27],[28], real datasets contain often outliers and noise. The data clusters are generally not well separated [15] and the corresponding probability density functions could be very dissimilar (effectives could be very different and density very dissemblable).
Non paramatric tools are classical to construct more robust methods. *Rank transformation* of data is one these ways [10],[12],[17].
We decided to use this transformation when fuzzifying data. You will note too that our fuzzification is deeply different from the concept of fuzzy rank statistic [13] despite the fact that the theoretical tools involved are quite the same. The rank transformation is introduced at the beginning of our method and is used in the computation of the membership function of each fuzzy subset $\tilde{x}_i$.

We suppose that we dispose of a dissimilarty measure on the set on Ω. Let δ denote this dissimilarity measure. Therefore, we can consider the induced weak orders family $(\precsim_i)_{i=1..n}$ such that $\forall i \in [1..n], \precsim_i$ is defined by :

$$\forall j,k \in [1..n], x_j \precsim_i x_k \Leftrightarrow \delta(x_i,x_j) \leq \delta(x_i,x_k)$$

Note : a relation on a set is called weak order if it is reflexive and transitive.

We can then obtain from this weak orders family the rank table $(R(i,j))_{i,j=1..n}$ defined by :

$$R(i,j) = \sigma_j^{-1}(i)$$

where $(\sigma_i)_{i=1..n}$ is the permutation family on Ω such as :

$$\forall i \in [1,n], x_{\sigma_i(1)} \precsim_i x_{\sigma_i(2)} \precsim_i \ldots \precsim_i x_{\sigma_i(n)}$$

In other words, the i^{th} column of R contains the ranks of the elements of Ω when sorting them by ascending value of dissimilarity measure with x_i.

This operation could be seen as a *rank transformation* that is a common way to make data analysis procedures more robust. The rank transformation consists in replacing each quantitative value of a variable in a multivariate dataset by the rank of this value. This transformation is frequently used and classical techniques like PCA have corresponding induced "nonparametric" methods.
In our case, we consider that the dataset is the given dissimilarity table. Each element is a datum and each datum is considerd as a descriptor too.
Let we describe this point. Let $D = (d_{i,j})_{i=1..n,j=1..n}$ denote this dissimilarity table of the dataset $(x_i)_{i=1..n}$. The element of the i^{th} row and the j^{th} column, $d_{i,j}$, is the dissimilarity measure between x_i and x_j. Each datum x_j induces a set of n real values corresponding to the j^{th} column of D and is viewed as a

variable.
The rank transformation replaces each of these real values by its rank when the set $(d_{k,j}), k = 1..n$ is ordered.

The first advantage of rank transformation is that the resulting methods are less sensitive to outliers. In fact one abnormal value generates high distances that could disturb a procedure base on the least square for instance. The corresponding ranks contain however no extreme values.

The second interest is that we make no assumption (of normality for instance). Ranks are uniformaly distributed. This point is attrative by notably allowing to deal with initially non symetric distributions.

According to this rank transformation we choose to consider that each datum of the dataset Ω is a fuzzy set where fuzzy membership function is a discrete function of the ranks of the data. If we consider the fuzzy set associated to $x_i \in \Omega$, the degrees of membership to $\tilde{x_i}$ are defined by :

$$\mu_{\tilde{x_i}}(x_j) = g(\sigma_i(j)) = g(R_{j,i}) \tag{1}$$

where g is a discrete monotonically decreasing function defined on $[1, n]$ and such that $0 \leq g \leq 1$.

The subsequently particularity of the obtained function is that its evaluation depends on the ranks of data. Each fuzzy datum $\tilde{x_i}$ is defined by using the same function on the ranks. We call this function the *shadow* of the generated fuzzy subset. If we consider the classical representation of membership function, we can observe that the *shapes* of the fuzzy data are however different. The choice of the *shadow* and his characteristics are presented in the following.
The shadow of the membership functions has to be choiced at the beginning of the processus. The role of this function could be compared to the kernel function in functional estimation problems.

In our context, it can be viewed as a scoring function. For instance, the value of $g(R_{ji})$ can be considered as a score assigned by $\tilde{x}_i$ to x_j. This score, assigned by x_i to x_k, is also a transformation of the rank assigned by x_i to x_j. The value of this score is the degree of membership of x_j to $\tilde{x_i}$.
The constraints on the shadow function g are easy to explain and interpret. For a given fuzzy data $\tilde{x_i}$ we want that the higher the rank of an element x_j the smaller its score. In other words, if the rank of x_j is small its degree of membership to $\tilde{x_i}$ has to be small. It justifies why g is forced to be decreasing. Therefore, since the result of g applied to a rank values is a degree of membership, g has to take values in $[0, 1]$

Let g be the Gaussian function: $g(r) = e^{-\frac{(r-1)^2}{s^2}}$ (where $s \in \mathbb{R}$ is the standard deviation of the Gaussian function g). Experimentally, we can affirm that this

Gaussian shadow is a satisfying choice by default. Other choices are proposed in the discussion part.

In this section we have presented our particular data fuzzification procedure where each element of the dataset is represented as one fuzzy set. We will now introduce a new notion by representing the whole data set as a fuzzy set itself.

2.2 The dataset as a particular fuzzy set

In this section we define the whole data set as a fuzzy set computed using the precedent step. This data set can be viewed as the aggregation of the n fuzzy sets associated to data. After presenting the theoretical definition we propose an application, and the interpretability of the notions will be exposed.

We define now a membership function of the fuzzy dataset $\tilde{\Omega}$ (the so called fuzzified Ω). This function is evaluated on each datum x_i. Let us describe this function evaluated on the datum $x_i \in \Omega$. Let $(\mu_{\tilde{x_1}}(x_i), \mu_{\tilde{x_2}}(x_i), \ldots, \mu_{\tilde{x_n}}(x_i))$ and $w = (w_1, \ldots, w_n)$ with each $w_i \in [0, 1]$ and $\sum w_i = 1$. We consider an ordered weighted averaging operator (OWA) [41] of dimension n, F_w :

$$F_w : \mathbb{R}^n \to \mathbb{R} : (y_1, y_2, \ldots, y_n) \mapsto \sum_{j=1}^{n} w_j.y_j$$

and we define : $\forall i \in [1..n]$,

$$\mu_{\tilde{\Omega}}(x_i) = F_w\left((\mu_{\tilde{x_1}}(x_i), \mu_{\tilde{x_2}}(x_i), \ldots, \mu_{\tilde{x_n}}(x_i))\right) \tag{2}$$

$\mu_{\tilde{\Omega}}(x_i)$ is the degree of membership of x_i in the fuzzy set $\tilde{\Omega}$.

2.3 Degree of representativeness

We have defined a particular way of representing data. It consists in representing each datum as a fuzzy set, and the whole dataset as another fuzzy set. It allows to represents data *individualy* and *globally* using the relevantness of fuzzy set theory.

We can see the degree of membership to $\tilde{\Omega}$ as a notion of *representativeness* resulting from the given dissimilarity measure. We can say that the high the degree of membership of one datum to the fuzzy dataset, the more representative the considered datum. This notion of representativeness is relative to the dataset.

It leads us to define the degree of representativeness.

Definition 1 *The* degree of representativeness *DR of one datum in Ω is defined by :*

$$DR(x_i) = \mu_{\tilde{\Omega}}(x_i)$$

We have exposed bellow a new framework for fuzzifying data. This approach can be relevant in the context of data analysis and we present in this section an example of application for the problem of unsupervised data clustering. The use and the interest of the degree of representativeness will be exposed in section 3.

3 Illustrations

After exposing the theoretical concept of our contribution, we present different kind of use in the field of data analysis. The first illustration introduce a purpose of representativeness in the problem of location estimation of a distribution. The second one illustrate the interest of representativeness in clustering problem.

3.1 The *Best Representative* : A robust location estimator

Let us consider the problem of location estimation [26],[3]. The goal of location estimation is to extract one datum or one value from the data sample or data space, that reflects the "location" of the data sample. In other terms, location estimation consists in finding one observation that is as representative of the sample as possible. A classical method consists in computing one statistic from the sample [26],[3],[39]. This statistic can be viewed as the estimation of one parameter of the underlying probability distribution [26]. The well known *mean* and *median* are the probably the most used statistics to achieve it. The *median* statistic is a robust location estimator contrary to the *mean*. It means that in presence of outliers, the *mean* becomes often non significative contrary to *median* that remains efficient. The robustness [39] is an important characteristics that makes the mean statistic obsolete when working on real data.

3.1.1 The best representative of a sample

We propose to use our representativeness notion as an objective function to be optimized in order to determine the best representantive of a sample. Trivially, the best representative of a data sample is the element of the set that is owning the maximum degree of representativeness.

Thus we define the statistic we called *best representative* as follow :

Definition 2 *The best representative of a data sample* $\Omega = \{x_1, \ldots, x_n\}$ *is the element of* Ω *whose degree of representativeness* $RD(x_i)$ *is the highest :*

$$x_{BR} = \arg\max_{x_i \in \Omega} DR(x_i)$$

Let us illustrate the interest and the characteristics of our statistic in some simulations.

3.1.2 Examples

The first point we want to highlight is the robustness [33] our the *best representative* of a sample. To exhibit it, we simulate a one dimensional data sample composed by one population distributed according to a non-centred χ^2 law (with 4 freedom degrees) , and an uniform noise. The noise represent 75% of the total sample. The figure 1 shows the distribution of the sample and the values of the *best representative statistic* versus the *mean* and the *median* statistics. We observe that the best representative lies in the "real" population contrary to the mean statistic which value shifts to the noise. As it's well known, the median is sufficiently robust to "resists" to this kind of noise.

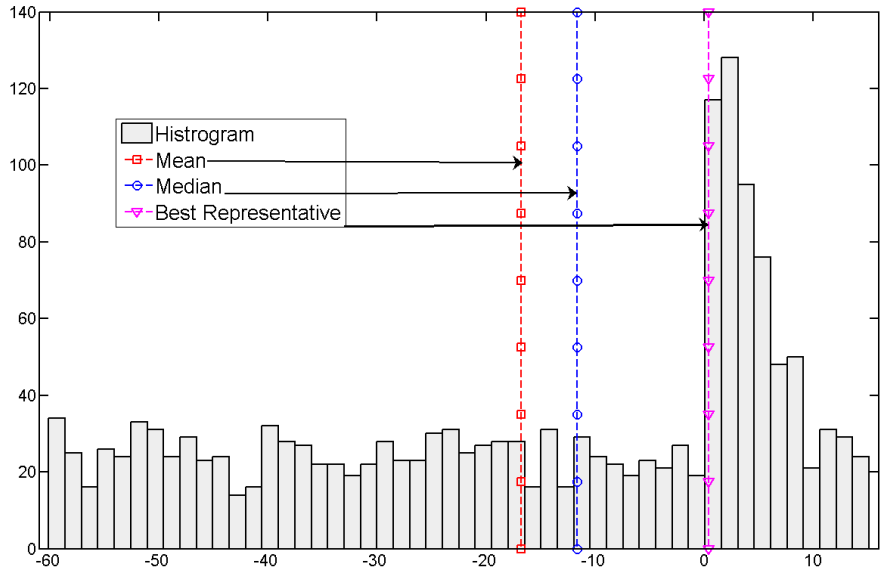


Figure 1: Best representative Vs Median Vs Mean - Data with 75% noise

The second characteristic of our statistics is the meaning it keeps when the sample is formed by several mixed subsamples. In fact, contrary to the mean and the median which only shows the central location of a sample, our statistic brings additional information. The *best representative* of a sample enhances the location information with information of representativeness. The figures 2 and 3 illustrate this point.

The figure 2 represents the distribution of a simulated random sample composed of two chi-square distributed subsamples ($\chi^2(4)$ and $\chi^2(24)$). It shows that the median and the mean of the sample give the central location of the whole data sample. They do not consider the two different subsamples contrary to the best representative that is located in the "middle" of one of the two subdistributions. The best representative corresponds to one element of the initial data sample that is as more representative as possible.

This result is illustrated in a second example (figure 3). In this third example, the sample is constitued of three subsamples that represent respectively 400, 200, and 100 elements of the sample. The median and the mean represent two elements of the minority subpopulation contrary to the best representative that is chosed in the majoritary sub-population.
This specificity is a real advantage in many cases. In fact, it permits to take account of the structure of the data sample. Classical statistics -that only traduce a central tendency of the sample- extract elements or compute values that are not representative in this given sample. If we consider for example a problem of consumer behavior, our statistic offers to determine the person which is the most representative in the sample of studied consumers. The median-people or the average-people are not necesseraly representative (median) or can be abstract (mean) (note that in the case of the *mean*, the results could be a "virtual average behavior" that corresponds to nobody in the sample).

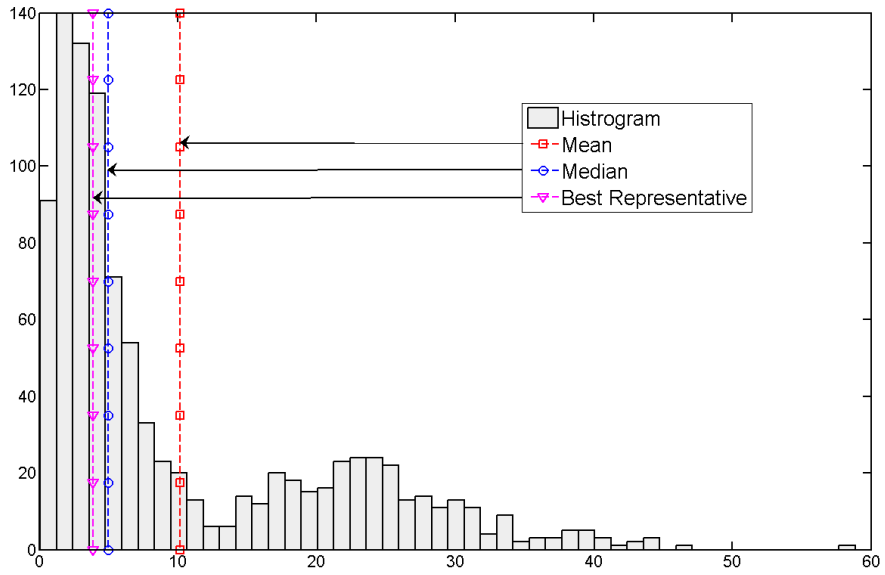


Figure 2: Best representative Vs Median Vs Mean - 2 subsamples

The numerical values of the statistics in these three examples are contained in the table 1.

3.2 A data clustering procedure based on representativeness

We will now describe another application of the representativeness degree in a sample. This application offers a clustering procedure based on the *degree of*

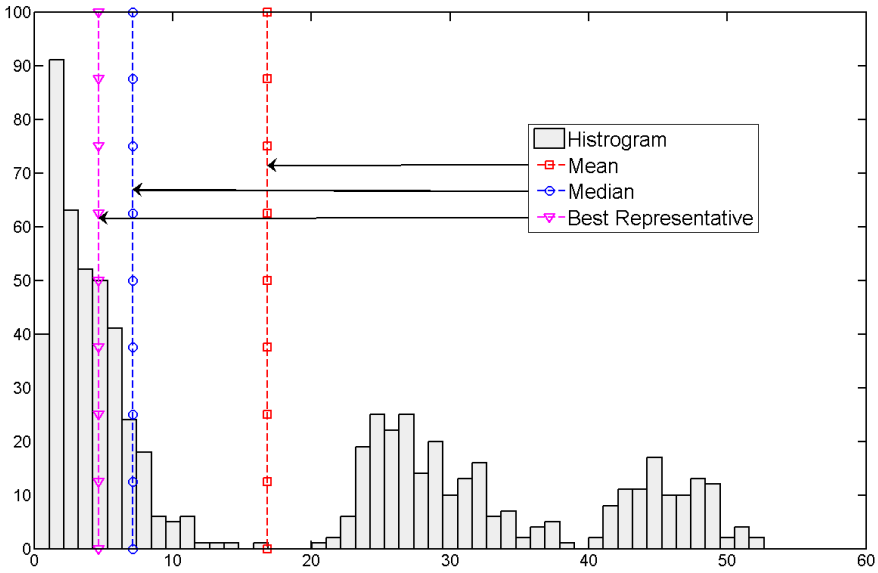


Figure 3: Best representative Vs Median Vs Mean - 3 subsamples

Statistic	Example of fig. 1	Example of fig. 2	Example of fig. 3
Mean	$-16,76$	$10,19$	$16,85$
Median	$-11,66$	$4,99$	$7,10$
Best representative	$0,44$	$3,89$	$4,67$

Table 1: Numerical values of the statistics in the three simulated examples

representativeness seen as an objective function to be optimized.

3.2.1 Principle

The principle lies on a classical approach in data clustering [22],[27]. In the cluster analysis field, we can find many algorithms based on the probability density function. The idea is to condider the underlying probability density function of the sample. Most of methods consists in estimating this density function and to associate each datum to one mode of the estimated function. This association induces a partition of the sample in clusters. More details about density based clustering algorithms can be find in [34],[9],[25],[11].

The idea of our application is to use the *degree of representativeness* of the data instead of the values of the estimated probability density function at these data.

This choice permits to keep the major advantage of these techniques which make no assumption on the shape of the clusters. That means for instance that it does not suppose that the clusters are spherical (it 's the case with the *k-means* algorithm for example).

The second advantage of our technique is its effectiveness with clusters which are distributed according to various densities and with clusters with different effectives, contrary to the classical techniques cited above.

The procedure we use is described as follow :

Algorithm 1 Representativeness Based Clustering Procedure

Require: The data sample $\{x_1, x_2, \ldots, x_n\}$, ε the radius of a sphere $S_\varepsilon(c)$ centered in c.
Ensure: The associated mode (AM) of each point of the data sample
  for all point x_i of the data sample **do**
    $AM(x_i) \leftarrow x_i$
    repeat
      $AM(x_i) \leftarrow \underset{x_k \in S_\varepsilon(AM(x_i))}{\arg\max} DR(x_k)$
    until stabilization
    return $AM(x_i)$ the mode associated to the point x_i
  end for

We will present our proposition in action in the following example.

3.2.2 Exemple

Let Ω be the two dimensional data sample whose graphical representation is presented in the figure 4 (Top left). After computing the representativeness of each

datum (represented in the figure 4(Top right)), we use the process described in the algorithm 1 to affect each datum to one mode of the representativeness function. At each iteration of the process, we shifts from the current point to the best representative of the neighborhood. So we obtain for each datum a path to the mode in the data sample. The figure 4(Bottom Left) represents the graph obtained by drawing these pathes on the representation of the representativeness. The so called graph is composed by two trees which define the data clusters.
In fact, by grouping data which have the same associated mode, we obtain the data clusters. Each final representative is the best representative of his induced cluster. The figure 4(Bottom right) represents the labelized data according to the obtaineed clusters.

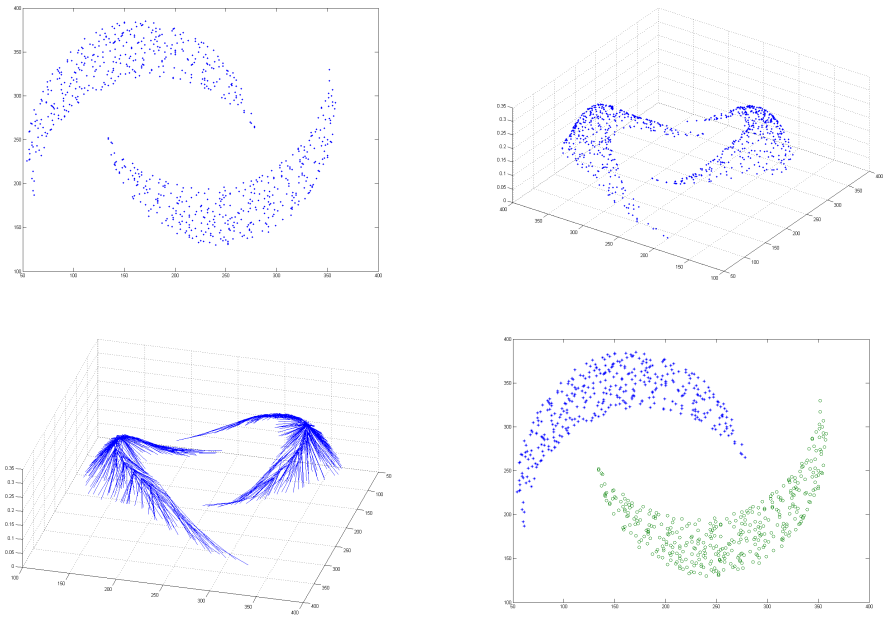


Figure 4: Illustration of using representtivity in a clustering process. Top Left : the two dimensional data sample. Top Right : 3D representation of the representativeness of each datum of the data sample (representativeness as elevation). Bottom Left : pathes obtained by iterative association to local best representative. Bottom Right : clusters obtained after labelization.

Despite the simplicity of the procedure, we constate immediately that our method permits to deal with the non linearity of the separation between clusters.

4 Discussion

We have presented below the theoretical aspects of our contribution, we discuss in this section some critical points and some choices we made.

4.1 Input data

The first point we have to discuss is the nature of the input data. Our method to represent data starts with using dissimilarity between data. The only essential precondition on the dataset is to have a way to compare data. This not-restrictive background allows our technique to be used with many kind of data and especially with multivariate data or non numerical data.
Therefore, as said in introduction of this paper, our method works on -and only on- the given data. The working set is the input dataspace. The computation of the representativeness is not depending on any visionary or unsupported assumptions about an underlying space of the data. Consequently, the use of this technique presents some practical interest. In our clustering application for instance, the representatives (the centers) of the clusters are extracted from the dataset. So they could be considered as "real" epitomes or "real" prototypes of the clusters, contrary to classical method that provides an imaginary mean representative (e.g. k-means).

The second point concerns the choice of the function that computes the degrees of membership to data as fuzzy sets from the ranks. We have called this function the shadow of the membership functions.

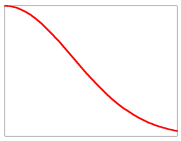
4.2 Shadow of membership function

The choice of the function is large but can be inspired or leaded from classical kernel functions or classical fuzzy membership functions (see for instance [8]) :

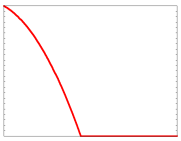
Let r be a rank value $R_{i,j}$,

- Gaussian shadow :
$$g(r) = e^{-\frac{(r-1)^2}{s^2}} \qquad s \in \mathbb{R}$$
- flate shadow :
$$f(r) = \begin{cases} 1 & \text{if } r \leq k \\ 0 & \text{if } r > k \end{cases} \qquad k \in \mathbb{N}^*$$
- Epanechnikov shadow :
$$k(x) = \begin{cases} k - r^2 & \text{if } r^2 \leq k \\ 0 & \text{if } r^2 > k \end{cases} \qquad k \in \mathbb{N}^*$$

Graphic illustrations of each of these shapes are presented in the table 2. The choice of the shadows is not so critical. It permits to give more flexibility to procedure for a better adaptability to the nature of data.


Gaussian shape


flate shape


Epanechnikov shape

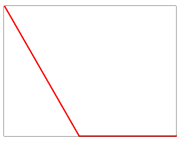

triangular shape

Table 2: Shapes examples

In our applications, we obtains good results with different shapes. The classical Gaussian function is usually a good choice and could thus be considered as an interessant choice by default. A different choice could be made according to the nature of the considered data of for particular applications.

4.3 Weights vector in OWA

The choice of the weights is very subsequent and could be guided by the input sample. We propose to design the weighting vector according to the trapezoid profil represented on the figure 5. Let us explain the way we choose. We consider that the first weights have to be null (I_1 on the figure 5). This constraint allows to ignore isolated data or very small group of data, when calculating the degree of membership of this data in the sample. In other words, it offers to treat the case of outliers. The second parameter permits to define the threshold of maximal contribution of one datum. So in this interval, the weights are maximum (I_2 on the Figure 5). Finally we consider that after a choosed threshold, the weights are null, i.e. the contribution of the datum is insignificant. Between these ranges, the weights are lineary calculated (I_3 on the Figure 5).

After discussing the different possible choices of the methodology, we present a different way of seeing our theoretical contribution. The idea is to expose another point of view on the defined notions by presenting these notions as objects in a different framework or context. In fact, our "fuzzy approach" can be expressed in a social choice theoretic context or in the framework of "preferences".

4.4 A different point of view

The degrees of membership to the sample fuzzy set can be interpreted as global scoring values affected to each datum by the set of all the others. This global score of one element is the result of the aggregation of the scores affected by each others. This point of view permits to make a link with the field of the social choice theory. In fact we can see the degrees of membership to each fuzzy data as the expression of individual preferences. On the other hand, the degrees of membership to the fuzzy data sample is close to the notions of collective

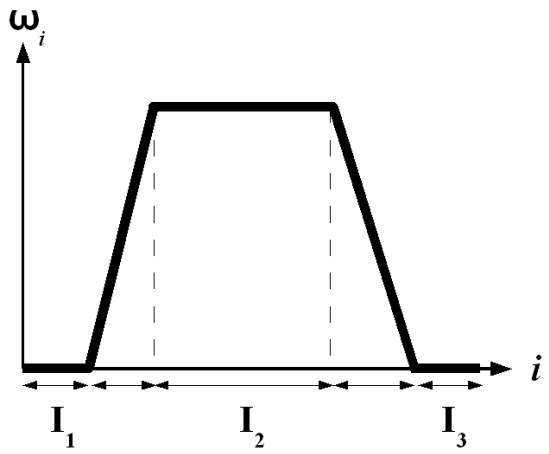


Figure 5: Construction of weight vectors for OWA procedure

preference.

5 Conclusion

We have presented in this paper the notion of representativeness in a data sample. The framework and the theoretical basis of this concept are the fuzzy set theory and data analysis. Our mathematical definition of the representativeness of data permits to address to the question "How one datum is representative in its data sample ?". The most immediate application of this approach is to construct a satistical location estimator which is robust and which keeps meaning and significance according to the underlying distribution. The use of representativeness in clustering problem is natural. By associating each datum to one "good representant", we create naturally a partition of the data set. Thus the representativeness notion provides a usefull tool for cluster analysis.

A Notations

Data	
Ω	the dataset
x_i	data ($\Omega = \{x_1, .., x_n\}$)
n	number of data ($n = Card(\Omega)$)
Rank transformation	
$\precsim_i$	weak order induced by the dissimilarity with x_i
$R_{i,j}$	Rank of x_j according to $\precsim_i$
Fuzzy data	
$\tilde{x_i}$	fuzzyfied data (each $\tilde{x_i}$ is a fuzzy set)
$\mu_{\tilde{x_i}}(x_j)$	membership degree of x_i to $\tilde{x_i}$
Agregation	
$w = (w_1, \ldots, w_n)$	weights vector used in OWA
F_w	OWA operator
Fuzzy dataset	
$\tilde{\Omega}$	fuzzyfied dataset
$\mu_{\tilde{\Omega}}(x_j)$	membership degree of x_j to $\tilde{\Omega}$

Table 3: Notations

References

[1] K. E. Barner, Y. Nie, and W. An. Fuzzy ordering theory and its use in filter generalization. *Journal on Applied Signal Processing*, vol. 4, pp. 206–218, 2001.

[2] V. Barnett. The ordering of multivariate data. *Journal of the Royal Statistical Society, Series A (General)*, vol. 139, no. 3, pp. 318–355, 1976.

[3] R. Beran. Robust location estimates. *The Annals of Statistics*, vol.5, no.3, pp.431–444, 1977.

[4] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

[5] D. G. Bonett. Approximate confidence interval for standard deviation of nonnormal distributions. *Computational Statistics and Data Analysis*, vol. 50, pp. 775–782, 2006.

[6] D. D. Boos. Comparing k populations with linear rank statistics. *Journal of the American Statistical Association*, vol. 81, no. 396, pp. 1018–1025, 1986.

[7] B. Bouchon-Meunier. *La logique floue*. PUF, 1993.

[8] Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, pp. 790–799, 1995.

[9] D. Comaniciu and P. Meer. Distribution free decomposition of multivariate data. *Pattern Analysis & Applications*, vol. 2, pp. 22–30, 1999.

[10] W. J. Conover and R. L. Iman. Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, vol. 35, no. 3, pp. 124–129, August 1981.

[11] J. Cutrona, N. Bonnet, and M. Herbin. A new fuzzy clustering technique based on pdf estimation. In *Information Processing and Management of Uncertainty*, pp. 225–232, Annecy, France, 2002.

[12] H. A. David and H. N. Nagaraja. *Order Statistics*. Wiley, 3 edition, 2003.

[13] T. Denoeux, M.-H. Masson, and P.-A. Hébert. Nonparametric rank-based statistics and significance tests for fuzzy data. *Fuzzy Sets and Systems*, vol. 153, pp. 1–28, 2005.

[14] J. J. Droesbeke and J. Fine. *Infŕence Non Paramétrique*. Ellipse, Paris, 1996.

[15] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, vol. 3, pp. 32–57, 1973.

[16] A. Flaig, K. E. Barner, and G. R. Arce. Fuzzy ranking: theory and applications. *Signal Processing*, vol. 80, pp. 1017–1036, 2000.

[17] M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937.

[18] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[19] J. Galambos. Order statistics of samples from multivariate distributions. *Journal of the American Statistical Association*, vol. 70, no. 351, pp. 674–680, September 1975.

[20] S. Guha, R. Rastogi, and K. Shim. Rock: a robust clustering algorithm for categorical attributes. In *In Proceedings of the 15th International Conference on Data Engineering*, pp. 512–521, 1999.

[21] V. Ha and P. Haddawy. Similarity of personal preferences: theoretical foundations and emprirical analysis. *Artificial Intelligence*, vol. 146, no. 2, pp. 149–173, 2003.

[22] J. A. Hartigan. *Clustering Algorithms*. Wiley, New York, 1975.

[23] T. C. Headrick and O. Rotou. An investigation of the rank transformation in multiple regression. *Computational Statistics and Data Analysis*, vol. 38, pp. 203–215, 2001.

[24] P.-H. Hébert, T. Denoeux, and M. Masson. Fuzzy rank correlation between fuzzy numbers. In *International Fuzzy Systems Association World Congress, IFSA*, pp. 175–186, 2003.

[25] M. Herbin, P. Vautrot, and N. Bonnet. Estimation of the number of clusters and influence zones. *Pattern Recognition Letters*, vol. 22, pp. 1557–1562, 2001.

[26] P. J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.

[27] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[28] R. Krishnapuram and J. M. Keller. A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, vol. 1, no. 2, pp. 98–110, 1993.

[29] T. Y. Lin. A set theory for soft computing - a unified view of fuzzy sets via neighborhood. In *Proceedings of the Fifth IEEE International Conference on Fuzzy Systems*, vol. 2, pp. 1140–1146, 1996.

[30] H. Mansouri. Multifactor analysis of variance based on the aligned rank transform technique. *Computational Statistics and Data Analysis*, vol. 29, pp. 177–189, 1999.

[31] R. Mesiar. Fuzzy set approach to the utility, preference relations, and aggregation operators. *European Journal of Operational Research*, vol. 176, no. 1, pp. 414–422, 2007.

[32] F. Murtagh. A survey of recent advances in hierarchical clustering algorithms which use cluster centres. *Computer Journal*, vol. 26, pp. 354–359, 1984.

[33] P. J. Rousseeuw and A. M. LeRoy. *Robust Regression and Outlier Detection*. Wiley, 2003.

[34] J. Sander, M. Ester, H.-P. Kriegel, and X. Xu. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data Mining and Knowledge Discovery*, vol. 2, pp. 169–194, 1998.

[35] I. R. Savage. Contributions to the theory of rank order statistics: Computation rules for probabilities of rank orders. *The Annals of Mathematical Statistics*, vol. 31, no. 2, pp. 519–520, 1960.

[36] H. Scheffe and J. W. Tukey. Non-parametric estimation. i. validation of order statistics. *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 187–192, 1945.

[37] M. Schemper. Generalized rank transformations for tests of survival. *Biometrical Journal*, vol. 33, no. 1, pp. 73–79, 2007.

[38] A. Slaby. Expectation of rank statistics under setup of stochastic inequalities. *Journal of Statistical Planning and Inference*, vol. 136, pp. 311–319, 2006.

[39] G. E. Thomas. Use of the bootstrap in robust estimation of location. *The Statictician*, vol. 49, no. 1, pp. 63–77, 2000.

[40] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[41] R. R. Yager. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 18, no. 1, pp. 183–190, 1988.

[42] L. A. Zadeh. *Fuzzy sets and systems.* J. Fox Polytechnic Press, New York, 1965.