A numerical distance based on fuzzy partitions

Serge Guillaume[1] **Brigitte Charnomordic**[2] **Patrice Loisel**[2]

[1]Cemagref, UMR ITAP, BP 5095, 34196 Montpellier, France
[2]INRA, SupAgro, UMR MISTEA, 34060 Montpellier, France

Abstract

This work studies a new distance function which takes into account expert knowledge by making use of fuzzy partitions. It considers the symbolic distances between concepts and is equivalent to the Euclidean distance for regular partitions made of triangular membership functions. Its behaviour is investigated in comparison with that of the Euclidean distance and its interest is shown for clustering applications.

Keywords: Similarity, interpretable, expert knowledge, k-means, clustering

1. Introduction

An important step in most clustering or classification techniques is to select a distance measure, which will determine how the similarity of two elements is calculated: similarity between individuals, between individual and group or between groups. The distance choice is a key point which has a strong impact on clustering results, see for instance [1, 10].

It is not a common practice to introduce expert knowledge in distance measures, though it is sometimes done through data transformations. The objective of the present work is to use fuzzy partitions to express expert knowledge on data, and to define a new distance based on these fuzzy partitions. This distance will be applicable to numerical data, while taking expert knowledge in account. Therefore it will constitute a compromise between a symbolic space - associated to a semantics - and a numerical one.

The concept of distance between fuzzy sets (or its dual concept of similarity) appeared early in fuzzy set theory for comparing or ranking purposes. Many works address this question, among them [6, 4, 13, 7, 3]. Some proposals paid attention to the fulfillment of the triangle inequality [2]. But, to our knowledge, the distance between individuals within a fuzzy partition was little-studied.

There are many ways to define distances between individuals. the most common ones for numerical values are defined by the L^p norms:

$$\|x\|_p = (|x_1|^p + \ldots + |x_n|^p)^{\frac{1}{p}}$$

where x is a multidimensional vector $(x_1, x_2 \ldots x_n)$.

The classical case of the Euclidean norm is obtained for $p = 2$.

When variables are not quantitative measurements, but nominal or categorical variables, other distances are available, based for instance on Percent disagreement. Our proposal aims to define a distance valid for numerical measurements, and having a symbolic component related to the granularity of information. For that purpose it is natural to use fuzzy partitions. Indeed they establish a correspondence between the numerical universe and linguistic variables, generally used in approximate reasoning and rule based systems. To respect the semantics attached to the partition-related concepts, the distance between two non distinguishable elements must be equal to zero. This is the case for all elements within a given fuzzy set kernel. Furthermore, elements belonging to different concepts must have a distance greater than elements belonging to the same concept.

In Section 2, we first propose a univariate function that meets these requirements. It is designed by deforming the Euclidean distance according to a standardized fuzzy partition structure. The particular case of regular fuzzy partitions is then studied. Proof of distance properties and a general definition in the multivariate case are done in Section 3. An illustration is presented for a clustering case in Section 4, including a comparison of results with the Euclidean distance. Some conclusions and perspectives are discussed in Section 5.

2. The proposed distance in the univariate case

The proposal combines numerical and symbolic elements. It applies to data in the unit interval [0,1] and it relies upon standardized, also called strong fuzzy partitions (SFP)[8, 9],such as the one plotted in Figure 1.

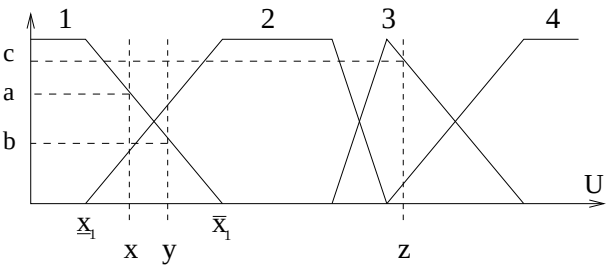

Figure 1: A standardized fuzzy partition

A SFP described by f membership functions (MF) on the universe U fulfills the following condition:

$$\forall x \in U, \qquad \sum_{i=1}^{f} \mu_i(x) = 1 \tag{1}$$

$\mu_i(x)$ is the membership degree of x to fuzzy set i.

A SFP is composed of two kinds of distinct areas: fuzzy set kernels, the sets of points which belong to a single MF, and overlapping zones between two concepts.

We consider partitions made up of linear triangular or trapezoidal MF.

The numerical part of the proposed distance allows to handle multiple membership in transition zones, while the symbolic one takes into account the granularity of the concepts associated to the fuzzy sets. Many possible combinations of numerical and symbolic elements may be studied, but it is not so easy to design one that will behave like a distance. To explain our choice, we first study the particular case of overlapping MF areas, before giving the general formula for the proposed distance.

2.1. Within overlapping areas

Definition 2.1 *Denote by I_i the overlapping area lying between the ith and the $(i+1)$th MF. Let $\underline{x}_i$ and $\overline{x}_i$ be the lower and upper bounds of I_i (see Figure 1 for I_1).*

In the overlapping area, the distance only consists of a numerical component. To obtain a smooth behavior, this numerical component d_n between two points is based on four membership degrees.

Definition 2.2 *Let x, y two points within the overlapping area I_i. The numerical component d_n is given by:*

$$\forall x < y \in I_i, d_n(x,y) = \mu_i(x)\mu_{i+1}(y) - \mu_{i+1}(x)\mu_i(y) \tag{2}$$

$$\forall x > y \in I_i, d_n(x,y) = d_n(y,x)$$

All membership degrees of x and y to MFs other than i, $i+1$ are equal to zero, due to the SFP structure.

From d_n definition we deduce the following properties:

Proposition 2.1 *The d_n value satisfies:*
(i) $d_n(x,y) > 0$ as $x \neq y$
(ii) for SFP, $d_n(x,y)$ becomes:

$$d_n(x,y) = \mu_i(x) - \mu_i(y) \text{ for } x < y. \tag{3}$$

Proof
(i) deduced from:
$\mu_i(x) > \mu_i(y),\ \mu_{i+1}(y) > \mu_{i+1}(x)$.
(ii) for SFP:

$$\forall x \in I_i, \qquad \mu_i(x) + \mu_{i+1}(x) = 1$$

□

The result is deduced using Equation 2. In that case, the numerical component is only defined by the membership degrees to a single MF, the first one in the partition order. For the example shown in Figure 1, $d_n(x,y) = a - b$.

Due to the linear nature of MF, we have:

$$\forall x \in I_i,\ \mu_i(x) = 1 - \frac{x - \underline{x}_i}{\overline{x}_i - \underline{x}_i}$$

Thus, given two points in the overlapping area, $d_n(x,y) = \dfrac{|y-x|}{\overline{x}_i - \underline{x}_i}$ is directly proportional to their Euclidean distance.

2.2. General case

When data points do not lie within the same overlapping area, as this is the case for points x and z in Figure 1, the formula is more complex. It is built to reflect the relative location within the MF each point belongs to, and also to take into account the number of MF between them. The former is a numerical component, and the latter is a symbolic one.

Notations: *Let us denote by M_x the number - within the partition - of the first MF for which the membership degree of the point x is strictly positive. $\mu_{M_x}(x)$ is the membership degree of x to M_x.*

Then we have $x \in I_i \Rightarrow M_x = i$

For the considered example in Figure 1, $M_z = 3$ and $M_x = 1$.

Definition 2.3 *Let x, z two points. The numerical component d_n is defined in a similar way to the previous case (see Equation 3):*

$$d_n(x,z) = \mu_{M_x}(x) - \mu_{M_z}(z)$$

The symbolic component d_s is defined as:

$$d_s(x,z) = M_z - M_x$$

The proposed univariate distance $d(x,z)$ is defined as the sum of the numerical and the symbolic components:

$$\forall x, z \in U,\ d(x,z) = d_n(x,z) + d_s(x,z) \tag{4}$$

For d_n, the underlying idea is a translation of point z in order to superpose the left bounds of the overlapping zones I_{M_x} and I_{M_z}.

The numerical component may be negative. This is illustrated on the figure as $a - c < 0$.

The symbolic component d_s compensates for the previous translation and takes into account the distance between fuzzy sets.

In order to compare future distances defined with respect to fuzzy partitions of different size, the sum of the components is normalized.

The proposed general formula for the univariate distance, valid whatever the location of x and y,

overlapping MF area, kernel area or no common MF, is then:

$$\forall x, y \in U,\ d_P^u(x,y) = \frac{|\mu_{M_x}(x) - \mu_{M_y}(y) + M_y - M_x|}{f-1} \tag{5}$$

where f is the number of fuzzy sets.

2.2.1. Equivalent definition from a function

The SFP-based proposed distance can also be designed using a function $Q(x)$.

Definition 2.4 *Within an overlapping zone, $Q(x)$ is defined as:*

$$\forall x \in I_i = [\underline{x}_i, \overline{x}_i], Q(x) = M_x - 1 + p_i(x)$$

where $p_i(x) = \dfrac{x - \underline{x}_i}{\overline{x}_i - \underline{x}_i}$ is the relative location of x in the interval.

Note that $M_x - 1$ is the number of MF located before the beginning of the overlapping area.

Given the relation between $\mu_i(x)$ and $p_i(x)$ we deduce:

Proposition 2.2 *The function Q satisfies:*

(i) $\forall x, Q(x) = M_x - \mu_{M_x}(x)$

In this form, $Q(x)$ is no longer restricted to overlapping zones but is generalized to the whole universe.

(ii) Q is a positive non decreasing function of x and is increasing in x in overlapping zones.

(iii) The univariate distance $d_P^u(x,y)$ defined in Equation 5 can then be written as:

$$\forall x, y \in U,\ d_P^u(x,y) = \frac{|Q(y) - Q(x)|}{f-1} \tag{6}$$

Proof

(i) from definition of $p_i(x)$ and $\mu_i(x)$.

(ii) from definition of $\mu_i(x)$. □

This formulation shows that the use of SFP partitions offers an elegant way to implement data transformation. It also makes it easier to check the properties of the proposed distance.

2.3. Particular case of regular SFP and Euclidian distance

Let us consider the special case of a triangle-shaped fuzzy partition whose vertices are regularly distributed in the universe. The quality of this type of partitions has been previously highlighted in the past, Pedrycz [14] pointed out that they are error free reconstruction for Sugeno type systems using centroid defuzzification:

$$\forall x \in U, \quad \psi[\phi(x)] = x$$

where ϕ is the input space transformation and ψ the output space one.

An example of regular SFP is given in Figure 2.

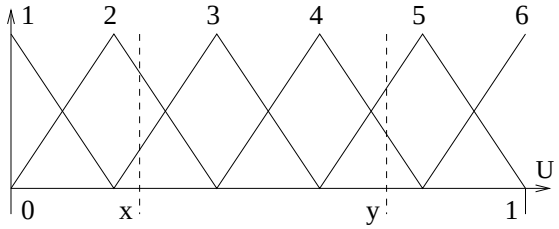

Figure 2: A regular strong fuzzy partition

As data are standardized in the unit interval, the regular distribution of the fuzzy set centers allows the simplification of Equation 6.

Proposition 2.3 *In the particular case of regular SFP, d_P^u is identical to the univariate Euclidean distance, remembering that data are in the unit interval:*

$$d_P^u(x,y) = |y - x|$$

Proof: for regular SFP:

$$\underline{x}_1 = 0 \text{ and } \forall i > 1,\ \underline{x}_i = \overline{x}_{i-1} = \frac{i-1}{f-1}$$

which implies $\overline{x}_i - \underline{x}_i = \dfrac{1}{f-1}$.

Therefore $p_i(x) = (f-1)x - (i-1)$ which implies that $Q(x) = (f-1)x$ and finally : $d_P^u(x,y) = |y-x|$. □

3. Properties and multidimensional case

3.1. Distance properties

A function d is a dissimilarity if:

$$\forall\ x, y \in U, \qquad \begin{cases} d(x,y) \geq 0 \\ d(x,x) = 0 \\ d(x,y) = d(y,x) \end{cases} \tag{7}$$

A dissimilarity is semi-proper if :

$$d(x,y) = 0 \Rightarrow \forall z \in U,\ d(x,z) = d(y,z)$$

A dissimilarity is proper if :

$$d(x,y) = 0 \quad \Rightarrow \quad x = y$$

A semi-distance is a dissimilarity which verifies the triangle inequality:

$$\forall\ x, y, z \in U, \qquad d(x,y) \leq d(x,z) + d(y,z)$$

A proper semi-distance is called a distance.

3.2. Checking properties

The proposed function d_P^u is a dissimilarity, properties mentioned in Equation 7 are trivially checked from Equation 6.

This is a semi-proper dissimilarity:

$$\begin{aligned}\forall x, y \in U,\ d_P^u(x,y) = 0 &\Rightarrow Q(x) = Q(y)\\ &\Rightarrow \forall z \in U,\ |Q(x) - Q(z)| = |Q(y) - Q(z)|\\ &\Rightarrow \forall z \in U,\ d_P^u(x,z) = d_P^u(y,z)\end{aligned}$$

In general, d_P^u is not a proper dissimilarity as the distance between two distinct, but not semantically distinguishable, elements is zero. The kernel of a given MF is a set of such counterexamples.

The triangle inequality fulfillment can be easily deduced from Equation 6:

$$\begin{aligned}\forall x, y \in U,\ (f-1)d_P^u(x,y) &= |Q(y) - Q(x)|\\ &= |Q(z) - Q(x) + Q(y) - Q(z)|\\ &\leq |Q(z) - Q(x)| + |Q(y) - Q(z)|\\ &\leq (f-1)(d_P^u(x,z) + d_P^u(z,y))\end{aligned}$$

The proposed function is thus a semi-proper dissimilarity which satisfies the triangle inequality, it is a semi-proper semi-distance. For the sake of simplicity, it is called a distance in the remainder of the paper.

3.3. Rank inversions

When the partition is not a regular triangle shaped fuzzy partition, the d_P^u distance is not any more the Euclidean distance, the deviations depending on the MF slope and kernel width.

This is illustrated in Figure 3. Let us consider three points x, y, z, with respective coordinates 0.2, 0.3, 0.5.

Therefore:

$$d_P^u(y,z) = 0.067 \quad < \quad d_P^u(x,y) = 0.133$$

while the Euclidean distances would yield:

$$d_{Euc}(y,z) = 0.2 \ > \quad d_{Euc}(x,y) = 0.1$$

3.4. Multidimensional distance

To obtain a multidimensional distance, the easiest way is to perform a Minkowski-like combination of the univariate distances. Let two multidimensional points $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$.

Their distance is defined as:

$$\forall x, y, \quad d_P(x,y) = \left[\sum_{j=1}^{n} d_j^k(x_j, y_j)\right]^{\frac{1}{k}} \qquad (8)$$

In Equation 8, each d_j is a univariate distance, either d_P^u as defined in Equation 5, which is based on a fuzzy partition, or a different one, for instance a univariate Euclidean distance. As d_P is a sum of semi-distance functions, it automatically inherits the properties of a semi-distance.

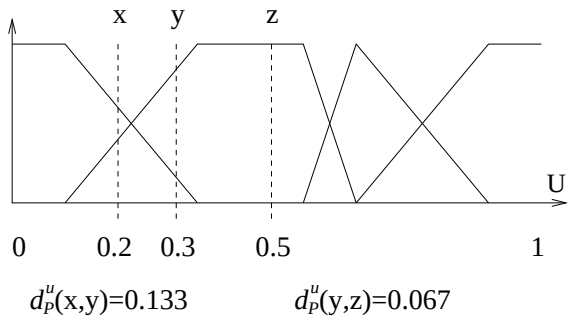

Figure 3: An example of inversion with the Euclidean distance

Note that, for k=2 and regular triangle shaped fuzzy partitions in all dimensions, Equation 8 yields the same result as the multidimensional Euclidean distance.

4. Clustering illustration

Data chosen to show the interest of the approach are taken from [11] and have already been used to illustrate clustering procedures [5]. Data describe the percentages of *water*, *protein*, *fat* and *lactose* in the milk of 22 mammals. They are given in Table 1.

	Water	Protein	Fat	Lactose	Ash
Bison	86.90	4.80	1.70	5.70	0.90
Buffalo	82.10	5.90	7.90	4.70	0.78
Camel	87.70	3.50	3.40	4.80	0.71
Cat	81.60	10.10	6.30	4.40	0.75
Deer	65.90	10.40	19.70	2.60	1.40
Dog	76.30	9.30	9.50	3.00	1.20
Dolphin	44.90	10.60	34.90	0.90	0.53
Donkey	90.30	1.70	1.40	6.20	0.40
Elephant	70.70	3.60	17.60	5.60	0.63
Fox	81.60	6.60	5.90	4.90	0.93
Guinea Pig	81.90	7.40	7.20	2.70	0.85
Hippo	90.40	0.60	4.50	4.40	0.10
Horse	90.10	2.60	1.00	6.90	0.35
Llama	86.50	3.90	3.20	5.60	0.80
Monkey	88.40	2.20	2.70	6.40	0.18
Mule	90.00	2.00	1.80	5.50	0.47
Orangutan	88.50	1.40	3.50	6.00	0.24
Pig	82.80	7.10	5.10	3.70	1.10
Rabbit	71.30	12.30	13.10	1.90	2.30
Rat	72.50	9.20	12.60	3.30	1.40
Reindeer	64.80	10.70	20.30	2.50	1.40
Seal	46.40	9.70	42.00	0.00	0.85
Sheep	82.00	5.60	6.40	4.70	0.91
Whale	64.80	11.10	21.20	1.60	0.85
Zebra	86.20	3.00	4.80	5.30	0.70

Table 1: Ingredients of mammal's milk

The aim of this experiment is to cluster the mammals, according to these five milk components, into a given number of groups, set to three. First a distance matrix between all the pairs of items is computed. Then, it is used as an input to the clustering algorithm. As the standard *k-means* does not give stable results with the Euclidean distance, a robust version, called *pam* (Partitioning around medoids)[12] in its R implementation[15], is used. The main difference between *pam* and *k-means* consists of the definition of the cluster centers: in the robust version, they are not computed as the mean but are necessarily one of the items, called a medoid.

4.1. Using the Euclidean distance

The results of the *pam* partitioning run on the multidimensional data are shown in Figure 4, which is a two dimensional plot. The three clusters are labelled E_1, E_2, E_3.

All observations are represented by points in the plot, using principal components analysis (PCA) to reduce the dimensions to the two first axes. An ellipse is drawn around each cluster. The first two components of the PCA explain 94.91% of the variability, and we will study the cluster composition on the first plane, also called principal plane. Of course some individuals may be closer or farther on the other factorial plans.

The cluster composition is a bit unexpected. The *dog* is included in the cluster of the sea mammals, while the *cat* and the *pig* are assigned to a different cluster.

A powerful indicator can be computed using the *silhouette* index [16]. To construct the silhouettes $S(i)$ for each item i, the following formula is used:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where $a(i)$ is the average dissimilarity of item i to all other items in the same cluster, and $b(i)$ is the minimum of average dissimilarity of item i to all items in other clusters. $b(i)$ can be seen as the dissimilarity between item i and its neighbor cluster, i.e., the nearest one to which it does not belong. The average silhouette width S_c for each cluster is simply the average of the $S(i)$ for all items in the *cth* cluster. Similarly the overall average silhouette width $\overline{S}$ is the average of the $S(i)$ for all items in the whole data set.

The silhouette index is based on cluster tightness and separation. It is followed from the formula that $-1 \leq S(i) \leq 1$. A value close to one indicates that the observation is correctly assigned to a group, a small value, even more so a negative one, witnesses a wrong assignment. The largest overall average silhouette indicates the best clustering.

Silhouette index values resulting from the Euclidean based clustering are given in the second row of Table 2, S_c for each cluster, and $\overline{S}$ for the overall

Euclidean	E_1	E_2	E_3	Overall
	0.27	0.28	0.55	0.35
SFP-based	P_1	P_2	P_3	Overall
	0.62	0.30	0.57	0.48

Table 2: Averaged silhouettes for each cluster and averaged overall value - Euclidean and SFP-based distances

averaged value. The same calculations will be done for the SFP-based distance.

4.2. Introducing expertise by the means of fuzzy partitions

The proposed distance allows the introduction of expert knowledge by the mean of fuzzy partitions. Two variables are considered, *Water* and *Fat*, the three other ones are handled using the Euclidean distance.

Figure 5 shows the design of the two groups, corresponding to *low* and *high* *Water* content.

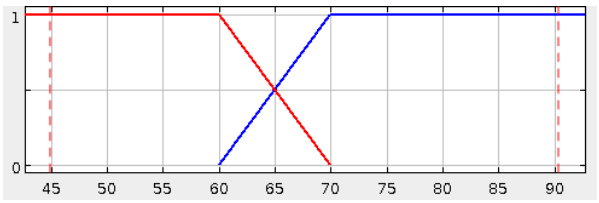

Figure 5: Fuzzy partition for *Water*

The *Fat* content variable has been partitioned into four groups as shown in Figure 6. This higher number of groups is motivated by the dispersion of the distribution. The ratio of the standard deviation to the mean is about 0.16 for *Water* content while it is higher than 1 for this variable.

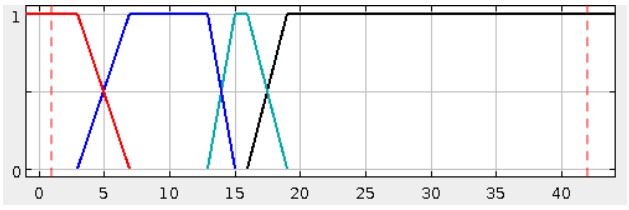

Figure 6: Fuzzy partition for *Fat*

Expertise consists, for instance, in considering that above 20%, *Fat* content is definitely high. It is a well known fact that marine mammal's milk has a much higher *Fat* content ($\geq 20\%$) than terrestrial mammal's milk, and that variations between 20% and 50% do not have that much importance.

The three new clusters are shown in Figure 7, and labelled P_1, P_2 and P_3. As could be expected, the cluster shapes are sensitive to the distance. Figure 7 can be compared with Figure 4. Changes can be

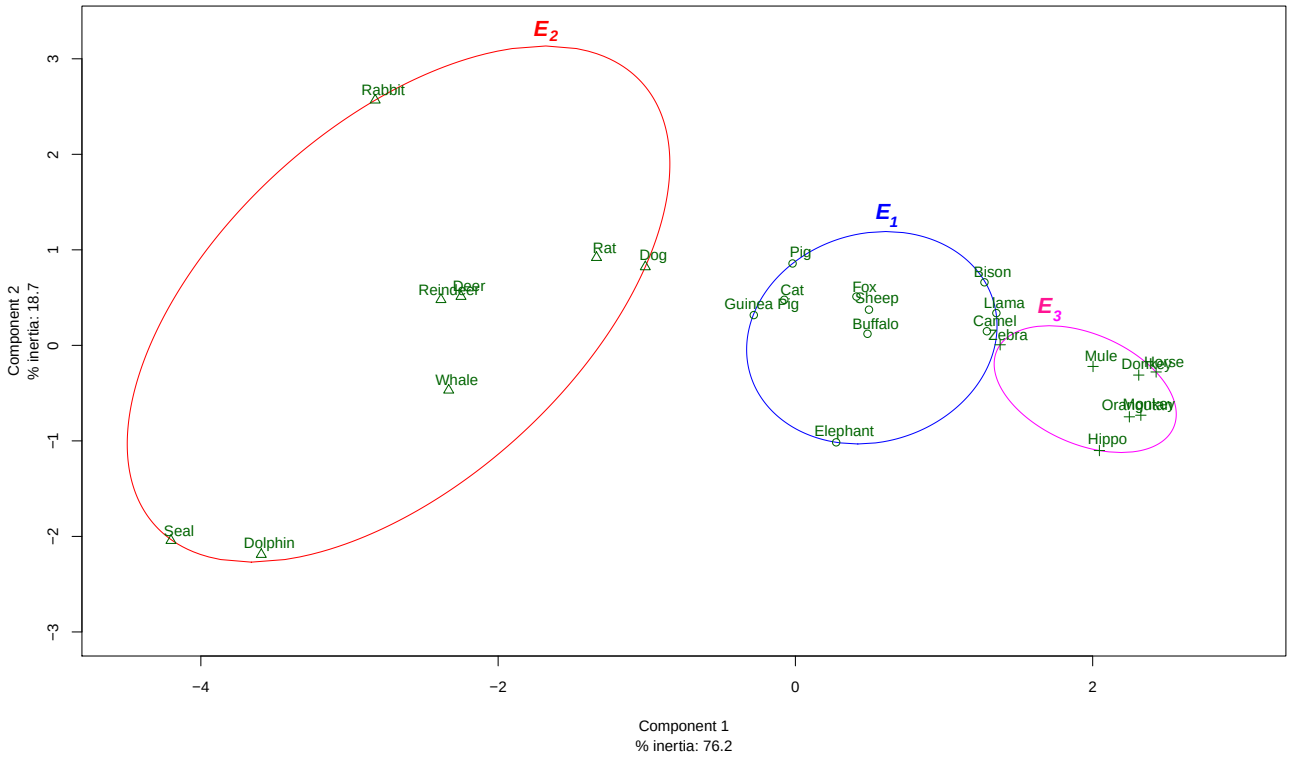

Figure 4: Clusters obtained with the Euclidean distance

noticed for all the clusters. The *pig*, the *dog* and the *cat* are now together in the central group. The boundary between the central and right clusters has also been modified. These two clusters are more neatly separated according to the plot. The *bison*, *camel* and *llama* are now in the group on the right, together with the *zebra*.

The new silhouette indices are given in the fourth row of Table 2. They also advocate for the SFP-based partitions: all S_c values are higher and the overall average $\overline{S}$ has been improved, 0.48 instead of 0.35.

5. Conclusion

In this paper we introduced a multivariate distance function that reflects the semantics of the fuzzy partitions defined on numerical domains.

This new multivariate distance is a combination of univariate distances. It operates just like the Euclidean distance, which it distorts according to the fuzzy partition structure. To define such a distance for a given dimension, a fuzzy partition must be specified for the corresponding linguistic variable.

The new univariate distance is a combination of symbolic and numerical terms.

The numerical term takes into account the multiple membership in transition zones from one concept to the next. The univariate distance between two points lying in the same transition zone is proportional to their Euclidean distance. When a point belongs to a transition zone, and another one is elsewhere, their distance combines a numerical part and a symbolic one.

The symbolic term factors in the distance between concepts, each concept being associated to a MF within the partition. All points within a given kernel are assumed to have a null distance. All concepts are considered as equidistant, independently of the Euclidean distance between kernel points.

The proposed function was shown to have the properties of a semi-proper semi-distance. As the multivariate distance performs a Minkowski-like combination of univariate distances, it can freely use different kinds of univariate distances. For instance, it can associate Euclidean distances in one dimension with symbolic ones in another one.

To illustrate the proposed distance, we applied it to a clustering case: mammal's milk. The results show the effect of the distortion of the Euclidean space on two variables among the five available ones. They highlight the cluster sensitivity to the distance choice. The new clusters are better separated and likely to be more interpretable.

Further work will be focused on the definition of such a distance based on fuzzy partitions more general than SFP. Targeted real world applications include merging processes in image analysis (region growing) and statistical procedures (hierarchi-

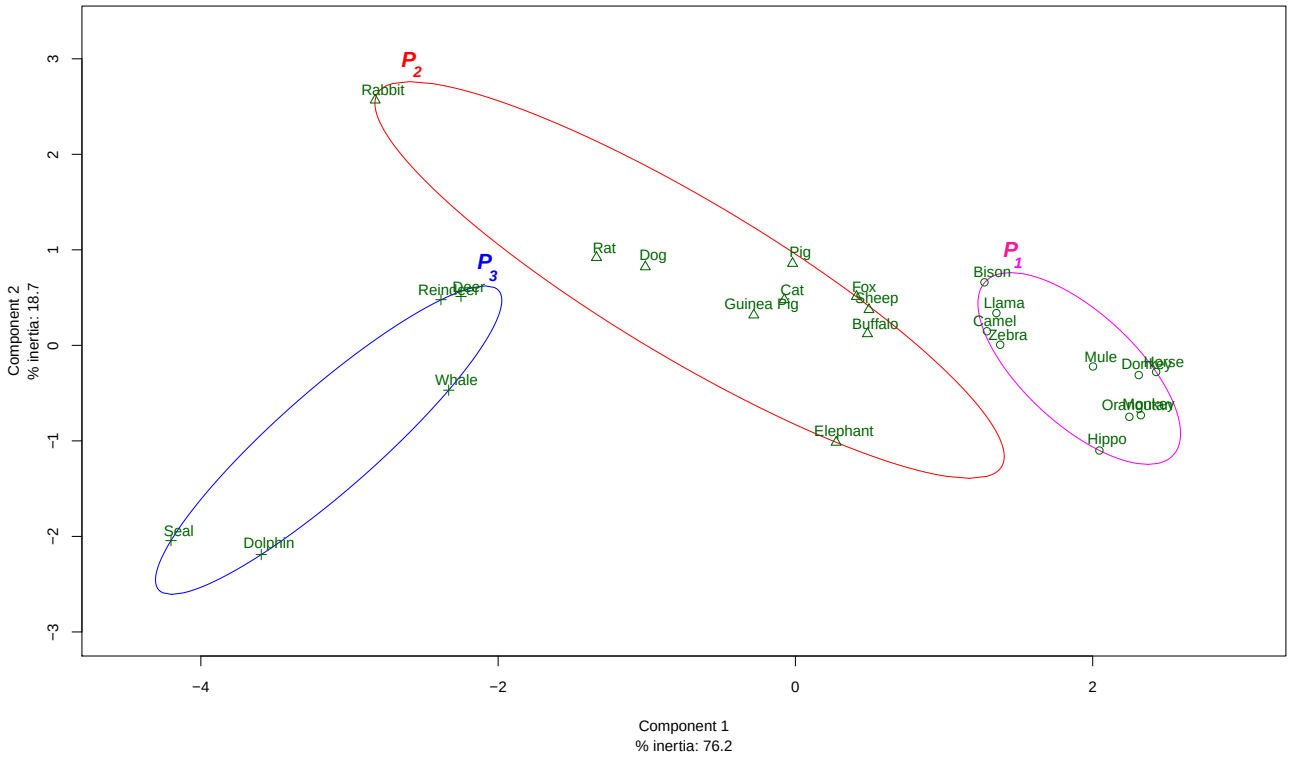

Figure 7: Clusters obtained with the new distance

cal clustering). Using the new distance in various application domains should draw attention to its potential for incorporating expert knowledge into learning methods.

References

[1] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Functions Algorithms, Plenum Press, New York, 1981.

[2] B. B. Chaudhuri, A. Rosenfeld, On a metric distance between fuzzy sets, Pattern Recognition Letters 17 (1996) 1157–1160.

[3] C. Coppola, T. Pacelli, Approximate distances, pointless geometry and incomplete information, Fuzzy Sets and Systems 157 (2006) 2371–2383.

[4] P. Diamond, P. Kloeden, Metric spaces of fuzzy sets, Fuzzy Sets and Systems 35 (1990) 241–249.

[5] W. J. Dixon, BMDP statistical software manual: to accompany the 1990 software release, BDMP (1990).

[6] D. Dubois, H. Prade, Fuzzy Sets and Systems: Theory and Applications, Academic Press, New York, 1980.

[7] J. Fan, W. Xie, Distance measure and induced fuzzy entropy, Fuzzy Sets and Systems 104 (1999) 305–314.

[8] S. Guillaume, Designing fuzzy inference systems from data: an interpretability-oriented review, IEEE Transactions on Fuzzy Systems 9 (3) (2001) 426–443.

[9] S. Guillaume, B. Charnomordic, Generating an interpretable family of fuzzy partitions, IEEE Transactions on Fuzzy Systems 12 (3) (2004) 324– 335.

[10] R. E. Hammah, J. H. Curran, On distance measures for the fuzzy k-means algorithm for joint data, Rock Mechanics and Rock Engineering 32 (1) (1999) 1–27.

[11] J. A. Hartigan, Clustering Algorithms, Wiley, 1975.

[12] L. Kaufman, P. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley Interscience, New York, 1990.

[13] R. Lowen, W. Peeters, Distance between fuzzy sets representing grey level images, Fuzzy Sets and Systems 99 (1998) 135–149.

[14] W. Pedrycz, Why triangular membership functions?, Fuzzy sets and Systems 64 (1) (1994) 21–30.

[15] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0 (2008).
URL http://www.R-project.org

[16] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53–65.