A distance metric suitable for fuzzy partitioning

Serge Guillaume
Cemagref
361 rue Jean-François Breton
34033 Montpellier - France

Brigitte Charnomordic, André Titli
INRA, LASB, 2 Place Viala, 34060 Montpellier, France
LAAS, CNRS, 7, avenue du Colonel Roche, 31400 Toulouse, France

Abstract— **In this paper, we address the problem to build up from data a fuzzy partition in an iterative procedure based on an original distance metric. Thanks to the used distances: internal distance within a fuzzy set, external distance between two fuzzy sets, distance between prototypes, the family of generated fuzzy partitions keeps the relevant properties of legibility and interpretability. A small and academic example, the iris data, demonstrates the efficiency of the proposed approach.**

Keywords—**Fuzzy partitioning, interpretability, distance, rule induction, learning**

I. Introduction

In order to manage complex systems, such as production processes in food industry, the decision maker has to take profit of different kinds of knowledge: expertise, local mathematical models and, probably, a lot of data ! We are here considering this last situation trying to structure the hidden knowledge in data, under the form of a fuzzy partition from which we want to build up a fuzzy rule base. This fuzzy rule base that will be used to develop approximate reasoning to take a decision must be interpretable, with a controlled complexity. Hence, the same properties are needed for the fuzzy partition we are looking for [1].

The importance of the concept of distance and the sensitivity of the results with respect to the choice of different distances has often been underlined in clustering [2], [3], but not in fuzzy partitioning. In fuzzy logic distances are usually defined between fuzzy sets [4], [5] for approximate reasoning. Some of the distances fulfill the triangle inequality [6], [7], other distances are pseudo metrics only [8]. The distance introduced by [9], [10] is close to human appreciation.

The goal of this paper is to propose the definition of internal and external distances of different data points in a fuzzy partition. These definitions are used to fuse adjacent fuzzy sets according to a given criterion. At each step of the iterative procedure we get an interpretable partition. The rest of the paper is organized as follows. In the section 2, we present the proposed distance metric introducing internal and external distances between data points, distance between prototypes. In the second part of the paper, we develop the hierarchical fuzzy partitioning procedure based on these distances. Due to the lack of space, we only illustrate the efficiency of the approach on a classical academic example with the iris data.

II. Proposed distance metric

In this section we first recall some basic properties of a distance, and we define some constraints on the fuzzy partition. Then we explain the basic concept of a pairwise distance metric compatible with multiple membership. We introduce the terms of internal distance, external distance, which are necessary for taking account of the fuzzy partition structure. Finally we give the definition of the proposed metric.

A. Distance properties

A function d is a dissimilarity if

$$\forall\, q,r \qquad \begin{cases} d(q,r) \geq 0 \\ d(q,q) = 0 \\ d(q,r) = d(r,q) \end{cases} \tag{1}$$

A dissimilarity is semi-proper if

$$d(q,r) = 0 \quad \Rightarrow \quad \forall\, s \qquad d(q,s) = d(r,s) \tag{2}$$

A dissimilarity is proper if

$$d(q,r) = 0 \quad \Rightarrow \quad q = r \tag{3}$$

A semi distance is a dissimilarity which verifies the triangle inequality

$$\forall\, q,r,s \qquad d(q,r) \leq d(q,s) + d(r,s) \tag{4}$$

A proper semi distance is called a distance.

B. Managing multiple membership

We limit our study to convex standardized fuzzy sets. The fuzzy sets have triangular or trapezoidal shape. They are labeled $1, 2, \ldots, m$ and they overlap so that the fuzzy partition is standardized as follows:

$$\begin{cases} \forall x \sum\limits_{f=1,2,\ldots,m} \mu^f(x) = 1 \\ \forall x \max\limits_{f=1,2,\ldots,m} \mu^f(x) \geq 0.5 \end{cases} \tag{5}$$

An illustration of the standardized fuzzy partition we are using is given on Figure 1. The definition (5) yields some useful properties for a fuzzy partition, as shown in a recent study [11]. However the proposed distance is general and can be applied to any kind of fuzzy partition.

Let us consider two data points with respective x_q^j, x_r^j coordinates in the jth dimension. Due to the fuzzification procedure, they can belong to several fuzzy sets with a non zero degree.

0-7803-7293-X/01/$17.00 © 2001 IEEE

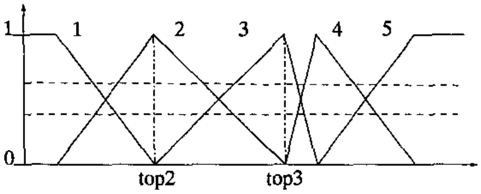


Fig. 1. A standardized fuzzy partition and the multiple membership bandwidth

When using a standardized partition, any data point has at most two non zero membership degrees. To alleviate the notations we will denote $\mu^f_q = \mu^f_j(x^j_q)$.

We introduce the terms of internal distance and external distance. Internal distance concerns partial membership of two points q, r to the same fuzzy set f, $\mu^f_q > 0$, $\mu^f_r > 0$.

External distance deals with their partial membership to two different fuzzy sets f and g, $\mu^f_q > 0$, $\mu^g_r > 0$, $f \neq g$.

We impose on any internal distance always to be less than any external distance. This fundamental restriction insures that the distance will reflect the partition structure and preserve the fuzzy set label semantic. Two points which mainly belong to the same fuzzy set will always be considered closer than others which mainly belong to distinct fuzzy sets.

C. Internal Distance

The membership degree complement $(1 - \mu^f_q)$ can be interpreted as the distance of x^j_q to the fuzzy set f. It measures the dissimilarity of x^j_q to the fuzzy set prototypes that delimit the kernel. Given two data points with x^j_q, x^j_r coordinates, we compute the internal distance by differencing the prototype similarities, which comes to differencing the membership degrees:

$$d_{int}(q,r) = |\mu^f_q - \mu^f_r|$$

Property (1) is trivial and equation (2) is easily checked. Counterexamples for property 3 are also easy to find. Many distinct data pairs have an identical membership degree, yielding a zero internal distance, as illustrated in figure 2.

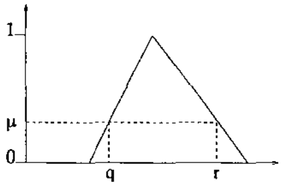


Fig. 2. Internal distance $d(q,r)$ equals zero

A (q,r,s) triplet is relevant of one of the following three cases for which Property (4) is to be checked.

1. Trivial case : identical membership for all three points: $\mu^f_q = \mu^f_r = \mu^f_s$.

2. Identical membership for two points: $\mu^f_q = \mu^f_r$ and $\mu^f_q \neq \mu^f_s$, with for instance $\mu^f_q > \mu^f_s$. The following inequalities are to be proven:

$$\begin{aligned} d_{int}(q,r) &\leq d_{int}(q,s) + d_{int}(r,s) \quad i.e. \quad 0 \leq 2(\mu^f_q - \mu^f_s) \\ d_{int}(q,s) &\leq d_{int}(q,r) + d_{int}(r,s) \quad i.e. \quad \mu^f_q - \mu^f_s \leq \mu^f_q - \mu^f_s \\ d_{int}(r,s) &\leq d_{int}(q,r) + d_{int}(q,s) \quad i.e. \quad \mu^f_q - \mu^f_s \leq \mu^f_q - \mu^f_s \end{aligned}$$

3. All membership degrees are distinct, for instance $\mu^f_q < \mu^f_r < \mu^f_s$. The inequalities to be checked are written as

$$\begin{aligned} \mu^f_r - \mu^f_q &\leq \mu^f_s - \mu^f_q + \mu^f_s - \mu^f_r \quad then \quad \mu^f_r \leq \mu^f_s \\ \mu^f_s - \mu^f_q &\leq \mu^f_r - \mu^f_q + \mu^f_s - \mu^f_r \quad then \quad \mu^f_s - \mu^f_q \leq \mu^f_s - \mu^f_q \\ \mu^f_s - \mu^f_r &\leq \mu^f_r - \mu^f_q + \mu^f_s - \mu^f_q \quad then \quad \mu^f_q \leq \mu^f_r \end{aligned}$$

In all cases the proof is straightforward therefore the proposed internal distance function d_{int} is a semi distance.

D. Prototype distance

Distances are made independent of measurement units by scaling data into the unit space. We propose two different definitions of the distance $d_{prot}(f,g)$ between the prototypes of fuzzy sets f and g. Recall that a prototype is a point x such as $\mu(x) = 1$.

1. A numerical distance: $d^{num}_{prot}(f,g) = \sqrt{(x_{k_f} - x_{k_g})^2}$. This definition corresponds to the kernel Euclidean distance. When using triangular fuzzy sets, the prototype is unique and corresponds to the coordinate of the triangle top, as shown in figure 1 for fuzzy sets 2 and 3. In the case of trapezoidal fuzzy sets, defined by their breakpoints a, b, c, d, all points in the interval $[b, c]$ are prototypes. The chosen prototype distance is the shortest one.

2. A more symbolic prototype distance: $d^{sym}_{prot}(f,g) = \dfrac{g-f}{m-1}$; where m is the partition size, f and g are the indices of the fuzzy sets sorted in ascending order when an order relation is meaningful on the fuzzy partition.

Within the partition illustrated on figure 1, the symbolic choice for the prototype distance makes the fuzzy set 3 at the same distance from 2 and 4, while the numerical choice puts it closer to 4. The symbolic distance is more faithful to the symbolic representation of the data.

Both definitions can easily be checked to fulfill conditions 3 and 4.

E. External distance

The external distance must take account of the point location within its reference fuzzy set, and of the relative fuzzy set location within the fuzzy partition, which implies combining the internal and the prototype distances.

We propose the following definition for the external distance between two points which belong to f and g:

$$d_{ext}(q,r) = |\mu^f_q - \mu^g_r| + d_{prot}(f,g) + D_c \tag{6}$$

where D_c is a constant correction factor, which ensures that the external distance is always greater than any internal distance.

Figure 3 illustrates the calculation of external distances on a (q,r,s) triplet. To simplify, let us consider only the membership to the three fuzzy sets labeled $(2,3,4)$. The external distances can be written as:

$$\begin{aligned} d_{ext}(q,r) &= \mu_r^3 - \mu_q^2 + d_{prot}(2,3) + D_c \\ d_{ext}(r,s) &= \mu_r^3 - \mu_s^4 + d_{prot}(3,4) + D_c \\ d_{ext}(q,s) &= \mu_q^2 - \mu_s^4 + d_{prot}(2,4) + D_c \end{aligned}$$

which proves the triangle inequality (4).

Note that the external distance reduces to the prototype distance plus the correction factor, when points q, r have internal identical membership degrees.

F. Distance combination and continuity

To manage multiple membership, the pairwise distance $d(q,r)$ is taken as a combination of the internal and external distances defined above, depending on the fuzzy sets for which μ_q and μ_r are non zero. In the most complicated case, both points have dual membership:

$$\begin{cases} \mu_q^f > 0, \mu_q^g > 0 \\ \mu_r^h > 0, \mu_r^l > 0 \end{cases} \tag{7}$$

Let us denote $d_{f,h}(q,r)$ the partial (q,r) distance that represents respective memberships to f and h. It is an internal distance if $f = h$, an external distance otherwise.

$d(q,r)$ results from the combination of four distances:

$$\begin{aligned} d(q,r) &= \frac{1}{\mu_q^f + \mu_q^g}\left(\mu_q^f * \frac{\mu_r^h * d_{f,h}(q,r) + \mu_r^l * d_{f,l}(q,r)}{\mu_r^h + \mu_r^l}\right. \\ &+ \left.\mu_q^g * \frac{\mu_r^h * d_{g,h}(q,r) + \mu_r^l * d_{g,l}(q,r)}{\mu_r^h + \mu_r^l}\right) \end{aligned} \tag{8}$$

For a standardized fuzzy partition, all denominators in the previous formula are equal to 1.

It is possible to use a simple way to manage multiple membership. Any point could be considered as mainly belonging to one fuzzy set, the one for which its degree is maximum, except those whose degrees fall into a bandwidth centered on 0.5, as shown on figure 1. It comes to consider as non significant all membership degrees below the bandwidth. Assume $\mu_q^f > 0$, $\mu_q^g > 0$. We deduce $max(\mu_q^f, \mu_q^g) \geq 0.5$, and $max(d_{int}(q,r)) \leq 0.5$. We therefore set $D_c = 0.5$ in Equation 6.

In the following formulae, all non significant degrees are set to zero.

Three cases are to be considered:

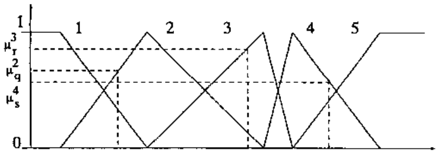


Fig. 3. An illustration of external distances

1. Each point belongs to one fuzzy set only. The distance is either purely internal or purely external:

$$d(q,r) = \begin{cases} d_{int}(q,r) = |\mu_q^f - \mu_r^g| & if\ f = g \\ d_{ext}(q,r) = |\mu_q^f - \mu_r^g| + d_{prot}(f,g) + 0.5 & if\ f \neq g \end{cases} \tag{9}$$

2. One point, say q significantly belongs to two fuzzy sets f and g. The other one belongs to h. The distance is evaluated as follows, assuming that $x_q^j < x_r^j$:

$$d(q,r) = \begin{cases} |\mu_q^g - \mu_r^g| + \frac{1}{2}\Big(d_{prot}(f,g) + 0.5\Big) & if\ g = h \\ d_{ext}(q,r) = |\mu_q^g - \mu_r^h| + d_{prot}(g,h) + 0.5 & if\ g \neq h \end{cases} \tag{10}$$

The simplification in 10 comes from assuming similar membership degrees $\mu_q^f \approx \mu_q^g$.

3. Both points have significant dual membership:

$$\begin{cases} \mu_q^f > 0, \mu_q^g > 0 \\ \mu_r^h > 0, \mu_r^l > 0 \end{cases} \tag{11}$$

Instead of combining all four distances, the economical way of managing continuity in the overlapping areas consists of keeping only the smallest one of the four. Equation 9 is thus used again and corresponds either to an internal distance or to an external distance.

The term $d(q,r)$ defined by (9) or (10), and in general by equation (8), has been shown to be a combination of semi distances. It is therefore a semi distance. Nevertheless to alleviate the notations we will refer to d as a distance.

III. Hierarchical fuzzy partitioning

Hierarchical Fuzzy Partitioning makes use of the proposed distance metric and generates a collection of univariate fuzzy partitions from a multidimensional training dataset. The data set is a collection of N multiple input-single output numerical data pairs $(x_k, y_k), k = 1, 2, \dots, N$ where x_k is the p dimensional input vector $x_k^1, x_k^2, \dots, x_k^p$ and y_k is the one-dimensional output vector. The method will be introduced in the next section, followed by an application to the iris Fisher data set.

A. Principle

A univariate fuzzy partition is composed of m fuzzy sets, the fth fuzzy set for the jth input variable being defined by its membership function $\Big(x, \mu_j^f(x)\Big)$. Denote m the fuzzy partition size.

A partition can be characterized by the standardized sum of pairwise distances over all the data points:

$$D_m = \frac{1}{N(N-1)} \sum_{q,r=1,2,\dots,N,\ q \neq r} d(q,r) \tag{12}$$

The procedure is carried independently over all dimensions. In each dimension it builds a family of fuzzy partitions as follows.

The initial fuzzy partition is determined by choosing M_j fuzzy sets according to the data sample distribution in the considered dimension, with $M_j \leq N$.

The family of fuzzy partitions is obtained using recursive fuzzy set merging so that at each step, the resulting partition is of size $m-1$, $m \geq 1$ and best satisfies a merging criterion. The final partition is composed of a single fuzzy set which covers the entire data range in the considered dimension. Merging is restricted to adjacent fuzzy sets, and seeks the best possible arrangement. Following merging, some external distances become internal distances, inducing a change on the D_m index. On figure 4, this is the case for all $d(q,r)$, $x_q^j \in [a,b]$, $x_r^j \in [c,d]$. The best merge at a given stage can be considered as the one that minimizes the variation of D_m. The underlying idea is to maintain as far as possible the homogeneity of the structure built at the previous stage. In some way the D_m index is analogous to the within class variance used in hierarchical clustering.

Each triangle fuzzy set f is defined by its breakpoints $left^f, top^f, right^f$. It is assigned a weight equal to its fuzzy cardinality $w^f = \sum \mu_j^f(x)$.

Merging two fuzzy sets labeled 2 and 3 is illustrated on figure 4. The resulting fuzzy set is labeled $2'$ and defined as follows:

$$\begin{cases} left^{2'} = top^1 \\ top^{2'} = \dfrac{w^2 * top^2 + w^3 * top^3}{w^2 + w^3} \\ right^{2'} = top^4 \end{cases}$$

The neighbouring fuzzy sets 1 and 4 are turned into $1'$ and $3'$. Their respective right and left breakpoints are modified so that the fuzzy partition is kept standardized.

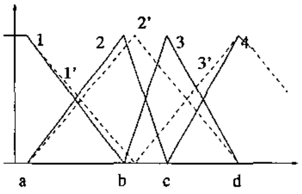


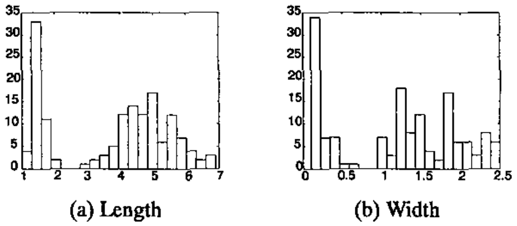
Fig. 4. Merging fuzzy sets 2 and 3 results in $2'$, $1 \Rightarrow 1'$, $4 \Rightarrow 3'$

B. Case study

The Hierarchical Fuzzy Partitioning method has been tested on the well known iris data, and especially the petal features, petal length and petal width. The corresponding histograms are plotted on figure 5. Remember that the iris are from three species Setosa, Virginica and Versicolor.

The numerical prototype distance is used. The results are reported in table I. The fuzzy set centers obtained with the proposed method, HFP, are compared with the ones found by the $k-means$ algorithm.

The petal length histogram clearly shows a two mode distribution. The second mode is issued from a combination of two iris species, very difficult to discriminate. For this reason we give the corresponding fuzzy partitions including two or three fuzzy sets.



(a) Length

(b) Width

Fig. 5. Histograms of *iris* petal features

	Petal width			Petal length				
	3 f. sets			3 f. sets			2 f. sets	
HFP	0.24	1.22	1.96	1.41	4.36	5.97	1.41	4.96
k-means	0.25	1.32	2.06	1.46	4.23	5.56	1.49	4.93

TABLE I

FUZZY SET CENTER COORDINATES FOR IRIS PETAL FEATURES

IV. CONCLUSION

In this paper we proposed a distance metric suitable for fuzzy partitioning, which we used for generating a family of univariate fuzzy partitions. For each partition of the family the fuzzy sets can be interpreted as linguistic labels. This is due to the standardized partition constraint maintained all through the procedure, and favored by the semantic attached to the distance used in the merging criterion. The key idea of differentiating the distances as internal and external distances makes it fit for reflecting the fuzzy set linguistic meaning. Such a partitioning can be useful as an input to rule induction methods, instead of classical grids or partitions derived from fuzzy clustering techniques.

The main benefit of the method is to leave the user with a possible choice of the fuzzy partition size. The final decision can be based on a priori knowledge. It can also be guided by validity criteria to be developed in further work.

REFERENCES

[1] Serge Guillaume, "Designing fuzzy inference systems from data: an interpretability-oriented review," *IEEE Transactions on Fuzzy Systems*, vol. 9 (3), pp. 426–443, June 2001.

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

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

[4] Didier Dubois and Henri Prade, *Fuzzy Sets and Systems: Theory and Applications*, Academic Press, New York, 1980.

[5] Jiulun Fan and Weixin Xie, "Distance measure and induced fuzzy entropy," *Fuzzy Sets and Systems*, vol. 104, pp. 305–314, 1999.

[6] B. B. Chaudhuri and A. Rosenfeld, "On a metric distance between fuzzy sets," *Pattern Recognition Letters*, vol. 17, pp. 1157–1160, 1996.

[7] B. B. Chaudhuri and A. Rosenfeld, "A modified hausdorff distance between fuzzy sets," *Information Sciences*, vol. 118, pp. 159–171, 1999.

[8] R. Lowen and W. Peeters, "Distance between fuzzy sets representing grey level images," *Fuzzy Sets and Systems*, vol. 99, pp. 135–149, 1998.

[9] Laszlo Koczy and Kaoru Hirota, "Ordering, distance and closeness of fuzzy sets," *Fuzzy Sets and Systems*, vol. 59, pp. 281–293, 1993.

[10] Pero Subasic and Kaoru Hirota, "Similarity rules and gradual rules for analogical interpolative reasoning with imprecise data," *Fuzzy Sets and Systems*, vol. 96, pp. 53–75, 1998.

[11] Jairo Espinosa and Joos Vandewalle, "Constructing fuzzy models with linguistic integrity from numerical data-afreli algorithm," *IEEE Transactions on Fuzzy Systems*, vol. 8 (5), pp. 591–600, October 2000.