Similarity, typicality and fuzzy prototypes for numerical data

Marie-Jeanne Lesot
Faculty of Computer Science, University of Magdeburg
lesot@iws.cs.uni-magdeburg.de

Abstract: In this paper, we consider the notions of similarity, typicality and prototype for numerical data, i.e. vectorial data belonging to R^p. We study the relationships between these notions and examine various possibilities to define similarity, to deduce typicality from similarity and prototypes from typicality, underlining the semantics and the interpretation of each considered method.

1. Introduction

The term *prototype* designs an element chosen to represent a group of data: it is an individual that characterises the group, summarises it and highlights its most important elements, facilitating its interpretation by the user. The prototype relies on a notion of typicality that models the fact that all members of a group are not equally representative or characteristic of it. Typicality has been studied at a cognitive and psychological level first by Rosch and Mervis[1]; they showed that the typicality of an element for a given category depends on two factors: its resemblance to the other members of the category and its differences to the members of other categories. The associated prototype then takes into account the common points of the group members, but also their distinctive features as compared to other categories. Building a prototype through typicality involves different ways of comparing data points: it depends both on similarity and dissimilarity measures.

In this paper, we consider these notions of comparison measures, typicality and prototype in the case of numerical data, i.e. vectorial data belonging to R^p, and we examine their relationships. We first consider various possibilities for the definition of dissimilarity and resemblance, highlighting their associated semantics. We then exploit them to define typicality degrees and fuzzy prototypes according to cognitive science principles. Lastly we consider other definitions of prototypes in the form of weighted means and interpret the weighting coefficients in the context of similarity and typicality. We show that most of them are to be interpreted as resemblance measures following local normalisation processes, suggesting ways to improve the typicality based prototype construction methods.

2. Comparison measures for numerical data

We first consider various possibilities for the definition of comparison measures for numerical data, i.e. resemblance and dissimilarity measures: usually dissimilarity is based on a distance function that is normalised in order to get a value in the interval [0, 1]; on the other hand, similarity is usually derived from scalar products or from dissimilarity measures through decreasing functions. In this section, we describe these methods and underline their semantics and properties that can enable a user to select the most appropriate.

2.1 Dissimilarity measures

[1] ROSCH E. and MERVIS C., (1975), *Family resemblance: studies of the internal structure of categories*, p. 573-605, Cognitive psychology, Vol. 7.

Distances: A classical way to define a dissimilarity measure consists in deriving it from a distance. Several definitions can be considered for this distance, the most common ones are described in table 1, their semantics and properties are discussed hereafter.

Table 1: Most commonly used distances and scalar products. p denotes the data dimension, (α_i) a vector of positive weighting coefficients and Σ the covariance matrix of the data.

Name	Distance	Scalar product
Euclidian	$d(x,y)=\sqrt{\sum_{i=1}^{p}(x_i-y_i)^2}$	$\langle x,y\rangle=\sum_{i=1}^{p}x_i y_i=x^T y$
Weighted Euclidian	$d(x,y)=\sqrt{\sum_{i=1}^{p}\alpha_i(x_i-y_i)^2}$	$\langle x,y\rangle=\sum_{i=1}^{p}\alpha_i x_i y_i$
Mahalanobis	$d(x,y)=\sqrt{(x-y)^t\Sigma^{-1}(x.y}$	$\langle x,y\rangle=x^T\Sigma^{-1}y$
Minkowski	$d_m(x,y)=\left(\sum_{i=1}^{d}\lvert x_i-y_i\rvert^m\right)^{\frac{1}{m}}$	

As compared to the Euclidian distance, its weighted variant offers the possibility to control the relative influence of the attributes and to rule their importance in the comparison. It is equivalent to performing a linear transformation of the data before using a Euclidian distance: each attribute is multiplied by the factor $\sqrt{\alpha_i}$. Thus it allows to normalise the attributes, which is indeed necessary when their values cover different scales: otherwise attributes taking high values dominate, the others having no influence in the comparison.

One can consider more general linear transformations by defining, for any symmetric definite positive matrix A, $d_A(x,y)=\sqrt{(x-y)^T A(x-y)}=d(A^{1/2}x, A^{1/2}y)$. The Mahalanobis distance belongs to this framework and takes as matrix A the inverse of the data covariance: it deduces the transformation from the statistical distribution of the data. If the covariance matrix is diagonal, the Mahalanobis distance corresponds to a weighted Euclidian distance with weights being the inverse of the standard deviation of each attribute; this is equivalent to normalising the attributes so that they all have mean 0 and variance 1. With a general Σ matrix, the Mahalanobis distance also takes into account correlations between attributes.

The Euclidian distance is a specific case of the two previous distances: it is a weighted distance with all coefficients equal 1, and a Mahalanobis distance with an identity covariance matrix. The level lines for the Euclidian distance are circles, in the weighted Euclidian case, they are ellipses parallel to the axes, and for the Mahalanobis distance, general ellipses. The selection of the most appropriate distance depends on the considered data, the statistical distribution of their attributes, and the importance the user wants to give to each attribute.

Besides, the Euclidian distance also is a Minkowski distance with $m=2$. Other often considered cases are the Manhattan distance ($m=1$) that has the advantage of being more robust than the Euclidian distance and the Chebychev distance ($m\to\infty$) that can also be written $d_\infty(x,y)=\max_{i=1..d}\lvert x_i-y_i\rvert$. One of the advantages of the Euclidian distance is to be a derivable function.

Normalisation: After a distance has been chosen depending on the distribution of the attributes and the desired properties (robustness, derivability e.g.), it must be normalised to

get a value in [0, 1] to define a dissimilarity[2]. Denoting d the distance to normalise, d_m and d_M its minimal and maximal values respectively, the simplest normalisation method consists in using the linear transformation

$$\eta(d) = \frac{d - d_m}{d_M - d_m} \tag{1}$$

Such a normalisation guarantees values between 0 and 1, these limits being obtained only for the extreme cases $d = d_m$ and $d = d_M$. The drawback of this approach is its sensitivity to outliers: the maximal distance can correspond to an aberrant point and be very large, disturbing the normalisation process.

Therefore one can use another transformation that overcomes this problem and offers additional properties, by defining the parameter $Z = (m, M)$ and

$$\eta_Z(d) = \min\left(\max\left(\frac{d-m}{M-m}, 0\right), 1\right) \tag{2}$$

Here, the extreme values are user-defined and not extracted from the data, which overcomes the data sensitivity problem. Moreover, this transformation provides a saturation property: it guarantees value 0 for any $d \leq m$ and value 1 for any $d \geq M$. Thus, the parameter m can be interpreted as a tolerance threshold: any distance lower than m leads to a zero dissimilarity, which means that distinct points can be considered as identical. The parameter M corresponds to the distance from which two data points are to be considered as totally dissimilar: the two points at maximal distance one from another are not the only ones to have dissimilarity 1; this is in particular interesting in the case of datasets containing aberrant points.

Thus normalisation makes it possible to modify the semantics of a dissimilarity through the definition of tolerance thresholds. We show further in section 4 how the normalisation parameters, which can be locally and not globally defined, can influence the semantics.

2.2 Similarity measures

Similarity measures have been formalised in the case of fuzzy data[2], i.e. data whose attributes are not numerical values but fuzzy subsets: this formal framework distinguishes between resemblance, satisfiability and inclusion, depending on the relationships between the two objects to be compared. In our case, none of the points is to be considered as a reference to which the other should be compared, they both have the same status; thus we consider resemblance measures. We will talk of similarity or resemblance without distinction.

We discuss here the two main approaches existing to define resemblance measures for numerical data: they can be deduced from scalar products, in particular kernel functions, or derived from dissimilarity measures through decreasing functions.

Scalar products: Scalar products can be written $\langle x, y \rangle = \|x\| \|y\| \cos(x, y)$ and take high values when the points are identical: they have the semantics of similarity, but must be normalised to get values in the interval [0, 1]. Table 1 mentions the most common scalar products, for which the previous discussion about attribute transformation holds.

Other semantically rich scalar products are defined in the kernel function framework, first introduced by Vapnik[3] and exploited in kernel learning methods (see e.g. Schölkopf and

[2] BOUCHON-MEUNIER B., RIFQI M. And BOTHOREL S., (1996), *Towards general measures of comparison of objects*, p. 143-153, Fuzzy sets and systems, vol. 84(2).

[3] VAPNIK V., (1995), *The nature of statistical learning theory*, New York: Springer.

Smola[4]). Indeed, they are associated to implicit nonlinear transformations of the data that enriches their representation: the polynomial kernel function for example is defined as

$$\rho_p(x,y) = (\langle x,y\rangle + l)^{\gamma} \quad (3)$$

In the case of 2D data for instance, $x=(x_1,x_2)$ and $y=(y_1,y_2)$, for $\gamma=2$ and $l=1$ one has

$$\rho_p(x.y) = (\langle x,y\rangle+1)^2 = (x_1y_1+x_2y_2+1)^2 = \langle \phi(x),\phi(y)\rangle$$

with $\phi(x) = (1, x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2)$. Thus this similarity is equivalent to considering a scalar product in a 5-dimensional space that takes into account quadratic correlations between the initial attributes. More generally, the polynomial kernel implicitly enriches the data with components defined as the monomials of degree inferior or equal to γ of the initial attributes. Thus it involves nonlinear correlations between the attributes without increasing the data dimensionality or the computational cost: the scalar product between the enriched data is computed using only their initial representation. Likewise the Gaussian function

$$\rho_g(x,y) = \exp\left(-\left(\frac{x-y}{\sigma}\right)^2\right) \quad (4)$$

corresponds to an infinite dimensional space for which the associated transformation function cannot be made explicit. Thus whereas the scalar products presented in table 1 can be seen as applying linear transformation of the data before comparing them, kernel functions offer the possibility to perform nonlinear transformations without increasing the computational costs. Therefore they constitute an interesting approach to the definition of resemblance measures.

Decreasing functions of dissimilarity: Besides, similarity measures can be derived from dissimilarity measures or distances, for instance as their complement to 1

$$\rho = 1-\eta_Z(d) \quad (5)$$

where η_Z is the normalisation function of eq. (2). It is to be noted that the normalisation parameters may be different from those defined for the dissimilarity itself: one can use as resemblance $\rho = 1-\eta_{Z_1}(d)$ and as dissimilarity $\delta = \eta_{Z_2}(d)$ with $Z_1 \neq Z_2$, leading to different threshold effects. One then considers that two points are totally dissimilar at a distance which differs from the distance at which their resemblance is null. We illustrate in section 3 the usefulness of such independent definitions.

One can also use smoother decreasing functions than this linear transformation to transform a dissimilarity measure to a resemblance; they provide means to influence the semantics of the resemblance. Table 2 presents three examples: the Laplace function, the generalised Gaussian function that corresponds to a Gaussian for $\gamma=2$, and the sigmoid or Fermi-Dirac[5] function. The latter requires a specific normalisation as indicated in the table and a saturation transformation $\max(f_{FD},0)$ to lead to interesting behaviours. These functions are illustrated for various values of their parameters on figures 1, 2 and 3 where the input variable represents a distance *d(x,y)* varying in the interval [0,1]. The Laplace and Gaussian functions are normalised using the function η_Z (cf. eq. (2)) with $m=f(1)$ and $M=f(0)$, the sigmoid function is normalised using its own procedure.

The first remark is that the global behaviour of the three families is quite similar, i.e. they offer the same variations, possibly for different parameters values. Graphs 1a, 2a and 3a show that in all 3 cases the parameter γ determines the decrease speed around $d=0.5$ and

[4] SCHÖLKOPF B. and SMOLA A., (2002), *Learning with kernels*, MIT Press

consequently the width of the plateaux for *d* lower or higher than the threshold 0.5. This corresponds to the *discrimination power* notion introduced by Rifqi et al.[5]: for high γ values, the Gaussian and Laplace curves have a steep decrease around 0.5, which implies that small differences in the input produce high differences in the output. On the contrary, the curve is very flat for small *d*, meaning that the function does not discriminate input values: different distances will produce approximately the same output. The same behaviour can be observed for the Fermi-Dirac functions for small γ values (see fig. 3a). When γ increases, the sigmoid function tends to a linear curve, whose discrimination power is equally spread on the whole

Table 2: Possible decreasing functions to define resemblance measures from dissimilarity measures.

Laplace	$f_l(x,y) = \dfrac{1}{1+\left(\dfrac{d(x,y)}{\sigma}\right)^{\gamma}}$	(6)
Generalised Gaussian	$f_{gg}(x,y) = \exp\left(-\left(\dfrac{d(x,y)}{\sigma}\right)^{\gamma}\right)$	(7)
Sigmoid function or Fermi-Dirac function[5]	$f_{FD}(x,y) = \dfrac{F(d(x,y)) - F(2\sigma)}{F(0) - F(2\sigma)}$ with $F(z) = \dfrac{1}{1+\exp\left(\dfrac{z-\sigma}{\gamma}\right)}$	(8)

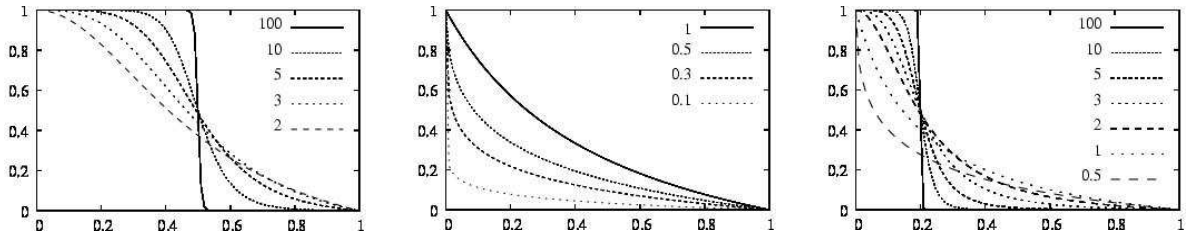


Figure 1: Laplace function after normalisation by the function η_Z , with $m = f_l(1)$ and $M = f_l(0)$ for various γ values: (a) $\sigma = 0.5, \gamma > 1$, (b) $\sigma = 0.5, \gamma \leq 1$, (c) $\sigma = 0.2$.

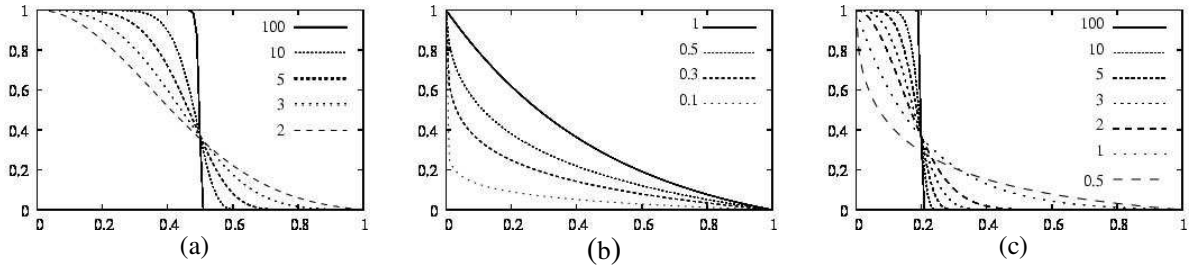
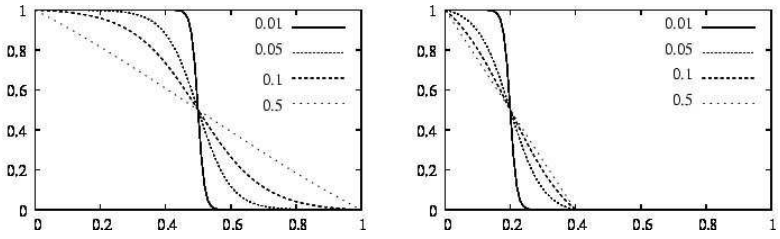


Figure 2: Generalised Gaussian function after normalisation by the function η_Z , with $m = f_{gg}(1)$ and $M = f_{gg}(0)$ for various γ values: (a) $\sigma = 0.5, \gamma > 1$, (b) $\sigma = 0.5, \gamma \leq 1$, (c) $\sigma = 0.2$

[5] RIFQI M., BERGER V. and BOUCHON-MEUNIER B., (2000), *Discrimination power of measures of comparison*, p. 189-196, Fuzzy sets and systems, Vol. 110.



(a) (b)

Figure 3: Fermi-Dirac function for several γ values: (a) $\sigma=0.5$, (b) $\sigma=0.2$

interval [0,1]. For the Gaussian and Laplace functions, one can observe a different comportment for $\gamma<1$, namely a high discrimination power for small input values (see fig. 1b and 2b): this behaviour is very strict as it implies that a distance, even very small, may decrease the resemblance value to a large extent. Thus the γ parameter enables the user to determine precisely the distribution of the discrimination power of the resemblance measure.

As for parameter σ, figures 1c, 2c and 3b show that it determines the position of the point with maximal discrimination power, i.e. the point where the derivative of the function is maximal (the latter is precisely located at $d=\sigma$ for the Laplace and Fermi-Dirac functions, a dependence to γ still exists in the Gaussian case, for which it is located at $((\gamma-1)/\gamma)^{\gamma}\sigma$). Therefore it also determines the threshold from which the output value is very small or zero.

Thus these functions make it possible to turn a dissimilarity measure to a similarity one, providing the user with parameters to control precisely its behaviour. It can be noted that the obtained functions can be reversed again to get dissimilarity measures: considering for instance $D=1-f_{gg}(d)$ leads to a smoother normalisation than the linear one (eq. (2)) and allows to determine the discrimination power of the dissimilarity measure.

3. Typicality degrees and fuzzy prototypes

The previous section discussed various possibilities for comparison functions, underlining their respective semantics: some modify the data representation, explicitly or not, in a linear or nonlinear way, others offer a fine-tuned control of the measure behaviour, in particular its discrimination power. In this section we illustrate the use of resemblance and dissimilarity for the definition of typicality degrees and fuzzy prototypes and underline the relationships between these notions.

Typicality degrees: Cognitive science works, initiated by Rosch and Mervis[1], have shown that all members of a category are not equivalent: some are more representative, or typical, than others. They showed further that a point is typical if it is similar to the other members of the group and distinct from members of other categories. This principle can be illustrated with the mammal category: whereas a dog can be considered as a typical example, a platypus is atypical because it does not resemble enough other mammals and a whale is atypical because it is not distinct enough from members of the fish category. This implies that typicality cannot be reduced to a resemblance notion: it is distinct from a simple similarity to the group centre, it also involves a dissimilarity notion.

Rifqi[6] proposed a method to implements these principles: it computes, for each point, its *internal resemblance*, i.e. its average resemblance to the other points of the group, and its

[6] RIFQI M., (1996), *Constructing prototypes from large databases*, p.301-306, Proc. of IPMU'96.

external dissimilarity, i.e. its average dissimilarity to points belonging to other categories. The typicality degree is then the aggregation of these two quantities.

Formally, let's denote $X=\{x_i, i=1..n\}$ the dataset, C_r , $r=1..c$ the categories, ρ and δ a resemblance and a dissimilarity measure respectively, as defined in the previous section. Then for point x_i and category r , the internal resemblance $R_r(x_i)$, the external dissimilarity $D_r(x_i)$ and the typicality degree t_{ir} are defined as

$$R_r(x_i) = \frac{1}{|C_r|}\sum_{y\in C_r} \rho(x_i, y) \qquad D_r(x_i) = \frac{1}{|X/C_r|}\sum_{z\notin C_r} \delta(x_i, z) \tag{9}$$

$$t_{ir} = \begin{cases} \phi(R_r(x_i), D_r(x_i)) & \text{if } x_i \in C_r \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where ϕ is an aggregation operator, such as the weighted mean or the symmetric sum e.g..

Fuzzy prototypes: The prototype is then defined as the aggregation of the most typical data: it highlights the common points of the category members but also their discriminative features. Several aggregation methods exist: Rifqi[6] considers the case of fuzzy data, the aggregation thus applies to fuzzy subsets and can be performed using classical fuzzy aggregation operators. In the case of numerical data, a solution may consist in defining the prototype as a weighted mean, using the typicality degrees as weighting coefficients, as $w_r = \left(\sum_i t_{ir} x_i\right) / \sum_i t_{ir}$: it indeed constitutes a way to aggregate numerical values. Yet it seems that a prototype should not be a single numerical value, but could be better modelled as a fuzzy set, even in the case of numerical values: prototypes correspond to imprecise notions that have vague boundaries; fuzzy subsets offer the flexibility to model these properties. Therefore Lesot et al.[7] proposed to define the prototype of numerical data as the fuzzy set whose kernel contains points having typicality higher than a threshold (e.g. 0.9), and its support contains points with typicality higher than a second, smaller, threshold (e.g. 0.7).

Numerical example: We illustrate this methodology and the criteria choice on the iris database that contains the length and width of petals and sepals for 150 flowers belonging to three categories, *virginica*, *setosa* and *versicolor*. We consider two attributes (sepal and petal widths) to allow a graphical representation and we build a prototype for each category.

We simply choose to base the comparison on a Euclidean distance as attributes have a similar value scale. The distance is then normalised using the function of eq. (2) to define resemblance and dissimilarity, with two different thresholds: the resemblance only takes into account distances at an intra-group level, thus an interesting distance reference is the maximal intra-group distance, i.e. the maximal diameter of the groups to be characterised. The dissimilarity is taken into account at an inter-group level, thus an interesting distance reference is the diameter of the whole dataset; the threshold is defined as the data half-diameter so as to avoid that a single point couple have a dissimilarity of 1. Lastly we apply a generalised Gaussian (eq. (7)) to soften the decrease of dissimilarity and to define the similarity from the normalised distance, considering

$$\rho(x,y) = f_{gg}(\eta_{Z_1}(d(x,y))) \quad \delta(x,y) = 1 - f_{gg}(\eta_{Z_2}(d(x,y)))$$

[7] LESOT M.-J., MOUILLET L. and BOUCHON-MEUNIER B., (2004), *Fuzzy prototypes based on typicality degrees*, 8th Fuzzy Days, Dortmund, Germany

where $Z_1=(0,\max_r(diam(C_r)))$ and $Z_2=(0,1/2\,diam(X))$; the parameters of the Gaussian are $\sigma=0.5$ and $\gamma=2$. The typicality degree is then defined using the weighted mean aggregator, as $t_{ir}=0.6R_r(x_i)+0.4D_r(x_i)$.

Figure 4 shows the level lines of the membership functions corresponding to the prototypes of the three categories. They are approximately centred on the group averages but are not spherical and their limits take into account the external dissimilarity: it can be seen that the two upper groups have a bigger spread in the x-direction as in the y-direction because the latter shows more overlapping between the groups. The lowest group is also clearly influenced by the two others that restrain it in the y-direction. Among the points having low membership to the prototypes, two categories can be made: some are located in overlapping

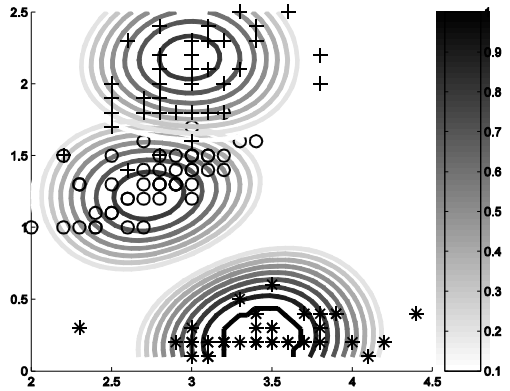


Fig. 4: Level lines of the membership functions for the fuzzy prototypes characterising the iris categories.

areas between categories, thus are rejected as having a too small external dissimilarity, others are too far away from the other members of the category, and thus have a too low internal resemblance. These prototypes incorporate information both about the common points and the distinctive features of the categories, and provide richer information than a numerical value due to their fuzzy properties.

4. Weighted means interpretation

In the previous section we considered a definition of typicality degrees and fuzzy prototypes according to cognitive principles. We discuss here other definitions of prototypes, in the form of weighted means, and compare the weighting coefficients to the notions of typicality and resemblance. We show that most of them correspond to internal resemblance with local normalisation but do not take into account an external dissimilarity component.

Average: As a reference, we consider the case of the arithmetic mean that is defined as

$$w_r=\frac{1}{|C_r|}\sum_{x_i\in C_r} x_i=\min_z \frac{1}{|C_r|}\sum_{x_i\in C_r}\|x_i-z\|^2$$

This corresponds, for point x_i and category *r*, to the weight

$$t_{ir}=\begin{cases}1 & \text{if } x_i\in C_r\\ 0 & \text{otherwise}\end{cases}$$

Such weights give the same influence to all points of the group and do not distinguish more or less representative points; they cannot be interpreted as typicality degrees. The second formulation above highlights the fact that the arithmetic mean minimises an internal dissimilarity, i.e. it is defined as the point which maximises an internal resemblance.

It is to be noted that the median is equivalent to the mean: it has the same definition, simply replacing the Euclidian distance by the Manhattan one, which makes it more robust.

***Most Typical Value*:** In order to build more significant representatives, Friedman et al.[8] define the *Most Typical Value* (MTV), as the fixed point of the equation

$$f(s) = \frac{\sum_{i=1}^{n} x_i f_i(|x_i - s|) m_i^{\lambda}}{\sum_{i=1}^{n} f_i(|x_i - s|) m_i^{\lambda}}$$

where m_i is the number of occurrences of the value x_i, λ is a user-defined parameter and f_i are decreasing functions that define local contexts and can differ for each point; they can for instance be the functions described in table 2. The resulting group representative can be seen as a weighted mean computed in an iterative process, where the weights are

$$t_{ir} = \begin{cases} m_i^{\lambda} f_i(|x_i - s|) & \text{if } x_i \in C_r \\ 0 & \text{otherwise} \end{cases}$$

Thus the influence of a point on a representative depends on its frequency and on its distance to the current estimate of the centre. The term $f_i(|x_i - s|)$ can be interpreted as an internal resemblance: it is equivalent to the quantity $R_r(x_i)$ defined in eq. (9), replacing the average similarity to the members of the group by the similarity to some average of the group.

The other element involved in the definition of t_{ir} does not correspond to an external dissimilarity, it is the frequency of the data points to the power λ : it models the expectation that a significant representative of the group should be all the closer to a point as the latter is frequent. In the typicality definition of eq. (10), each occurrence of a point is handled independently, which is equivalent to considering only the specific case $\lambda = 1$.

The MTV can be interpreted as a more relevant representative than the arithmetic mean relying on a notion of internal resemblance that takes into account the data frequency.

Fuzzy *c*-means: The fuzzy *c*-means is a clustering algorithm that can also be seen as computing a weighted average in an iterative process. The coefficients are u_{ir}^m where *m* is a user-defined parameter, and

$$u_{ir} = \frac{1}{\sum_{s=1}^{c} \left(\frac{\|x_i - w_r\|}{\|x_i - w_s\|} \right)^{\frac{2}{m-1}}} = \frac{1}{1 + \|x_i - w_r\|^{\gamma} \sum_{s \neq r} \frac{1}{\|x_i - w_s\|^{\gamma}}}$$

with $\gamma = 2/(m-1)$ and w_r , $r = 1..c$ the current estimates of the group representatives. An important difference to the previous coefficients is the fact that these weights never equal 0: a point influences the representatives of all categories and not only a single one. This provides the algorithm with interesting properties[9] but also leads to difficulties for the coefficient interpretation: if for instance two categories are considered, all points located on the median between the two centres have weight 0.5 for both groups, independently on their actual distance to the centres. This behaviour implies that the coefficients do not have the semantics

---

[8] FRIEDMAN M., MING M., and KANDEL A., (1995), *On the theory of typicality*, p. 127-143, International Journal of Uncertainty and Knowledge-Based Systems, Vol. 3.

[9] KLAWONN F., (2005), *Understanding the membership degrees in fuzzy clustering*, Proc. of the 29th Annual GfKl Conference, Springer-Verlag.

of typicality degrees nor of a resemblance measure, from both of which it is expected that they decrease with the overall distance. In fact these coefficients are interpreted as membership degrees, indicating the degree with which a data point belongs to each group, or sharing coefficients, indicating the extent to which it is shared between the groups.

The second formulation above indicates that u_{ir} is a Laplace transformation (cf. eq. (6)) with $\gamma = 2/(m-1)$ applied to the distance to the centre current estimation, which *a priori* gives it the interpretation of an internal resemblance. The difference comes from the definition of σ that normalises the distances: it is not defined globally, but varies for each point and each group, defined by $\sigma_i^{\gamma} = \left[\sum_{s \neq r} 1/\|x_i - w_s\|^{\gamma}\right]^{-1}$. Thus for a point x_i, $\|x_i - w_r\|$ is compared to its distance to the centres of other groups. This relative distance indicates the extent to which x_i is shared between the groups. It modifies the interpretation of the Laplace function as a similarity measure, which underlines the influence of normalisation on the global semantics.

We commented here on coefficients u_{ri}, whereas the group representatives are based on their value to the power of *m*. This transformation does not modify the semantics, it can be considered as an optimisation trick that distinguishes the fuzzy *c*-means from Gaussian mixture models and the EM algorithm[10].

Possibilistic *c*-means: The possibilistic *c*-means[11] is another clustering algorithm, it relies on weights that depend on parameters η_r indicating the diameter of each group: the latter can be deduced from *a priori* knowledge or determined using the fuzzy *c*-means in a preliminary step[11]. The weights are then defined as t_{ir} to the power *m*, where

$$t_{ir} = \frac{1}{1 + \left(\frac{\|x_i - w_r\|}{\eta_r} \right)^{\frac{2}{m-1}}}$$

They correspond to a Laplace transformation of a distance with exponent $\gamma = 2/(m-1)$ and a locally defined parameter σ. Contrary to the fuzzy *c*-means, the normalisation does not vary for each data point, but only depends on the group to be characterised and equals the diameter of the group. In the experiment with the iris dataset (cf. section 3) we considered a similar approach but we used a single parameter, the maximal value of these diameters, and not a different value for each category: the resemblance measure used here is more refined.

It is to be noted that, as in the fuzzy *c*-means case, these coefficients *a priori* never take zero values. Yet, for points close to other groups, $\|x_i - w_r\|$ is higher than η_r, thus t_{ir} takes small values or equals 0: in practical cases, points significantly contribute to a single centre.

Fuzzy possibilistic fuzzy *c*-means: Pal et al.[12] define a clustering algorithm that combines the characteristics of both fuzzy and possibilistic *c*-means: it relies on two weight distributions, one being identical to the fuzzy *c*-means distribution, the other being defined as

---

[10] DÖRING C., BORGELT C. and KRUSE R., (2004), *Fuzzy clustering of quantitative and qualitative data*, p. 84-89, Proc. of the Conf. of the North American Fuzzy Information Processing Society.

[11] KRISHNAPURAM R. and KELLER J., (1993), *A possibilistic approach to clustering*, p. 98-110, IEEE Transactions on fuzzy systems, Vol. 1.

[12] PAL N., PAL K. and BEZDEK J., (1997), *A mixed c-means clustering model*, p. 11-21, Proc. of the IEEE Int. Conf. On Fuzzy Systems.

$$t_{ir} = \frac{1}{\sum_{j=1}^{n} \left(\frac{\|x_i - w_r\|}{\|x_i - w_r\|} \right)^{\frac{2}{m-1}}}$$

to the power *m*. It is close to the quantity considered by the fuzzy *c*-means, but it differs by the normalisation: the latter is the same for all points in a group and is defined by $\sigma^{\gamma} = \left[\sum_j 1/\|x_j - w_r\|^{\gamma} \right]^{-1}$: this sum is over data points and not cluster centres. Thus it can be seen as an indirect measure of the spread of the cluster, i.e. it is more similar to a diameter. Thus this distribution is again to be interpreted as an internal resemblance which takes into account a local normalisation.

5. Conclusion

We considered the problem of similarity, typicality and fuzzy prototypes in the case of numerical data. We mentioned some of the various possibilities that exist to compare vectorial data and provide users with much more flexibility than the simple Euclidian distance. We illustrated the use of these measures for the definition of typicality degrees corresponding to psychological and cognitive studies. Lastly we examined some weight distributions used to define data representatives and compared them with the notions of internal resemblance and typicality. We showed that a highly influential criterion is the distance normalisation, which determines the semantics of the relative distance and thus that of the overall transformation.

This comparison highlights the fact that both approaches could benefit one from another: the existing weights suggest ways to enrich internal resemblances and possibly external dissimilarities, involving the data frequency for instance or using local normalisations such as the group diameters for each category. Reciprocally, one could perform clustering taking into account an external dissimilarity component and defining cluster centres that underline the distinctive features of clusters.